Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical

Khadim Drame

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

Ecole Doctorale Sociétés, Politique, Santé Publique

Spécialité : Informatique et Santé

Par Khadim DRAME

Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical

Sous la direction de : Roger SALAMON

Soutenue le 10 décembre 2014

Membres du jury :

M Guy Melançon	Pr, Université de Bordeaux	Président
M Moussa Lo	Pr, Université Gaston Berger	Rapporteur
M Pierre Zweigenbaum	DR CNRS, LIMSI	Rapporteur
Mme Nathalie Hernandez	MCF, Université Toulouse II	Examinateur
M Jean François Dartigues	PU-PH, Université de Bordeaux	Examinateur
M Gayo Diallo	MCF, Université de Bordeaux	Co-encadrant
Mme Fleur Mougin	MCF, Université de Bordeaux	Co-encadrant

Résumé

Ce travail vise à permettre un accès efficace à des informations pertinentes malgré le volume croissant des données disponibles au format électronique. Pour cela, nous avons étudié l'apport d'une ontologie au sein d'un système de recherche d'information (RI).

Nous avons tout d'abord décrit une méthodologie de construction d'ontologies. Ainsi, nous avons proposé une méthode mixte combinant des techniques de traitement automatique des langues pour extraire des connaissances à partir de textes et la réutilisation de ressources sémantiques existantes pour l'étape de conceptualisation. Nous avons par ailleurs développé une méthode d'alignement de termes français-anglais pour l'enrichissement terminologique de l'ontologie. L'application de notre méthodologie a permis de créer une ontologie bilingue de la maladie d'Alzheimer.

Ensuite, nous avons élaboré des algorithmes pour supporter la RI sémantique guidée par une ontologie. Les concepts issus d'une ontologie ont été utilisés pour décrire automatiquement les documents mais aussi pour reformuler les requêtes. Nous nous sommes intéressés à : 1) l'identification de concepts représentatifs dans des corpus, 2) leur désambiguïsation, 3), leur pondération selon le modèle vectoriel, adapté aux concepts et 4) l'expansion de requêtes. Ces propositions ont permis de mettre en œuvre un portail de RI sémantique dédié à la maladie d'Alzheimer. Par ailleurs, le contenu des documents à indexer n'étant pas toujours accessible dans leur ensemble, nous avons exploité des informations incomplètes pour déterminer les concepts pertinents permettant malgré tout de décrire les documents. Pour cela, nous avons proposé deux méthodes de classification de documents issus d'un large corpus, l'une basée sur l'algorithme des k plus proches voisins et l'autre sur l'analyse sémantique explicite. Ces méthodes ont été évaluées sur de larges collections de documents biomédicaux fournies lors d'un challenge international.

Mots clés : construction d'ontologie, réutilisation de RTO, recherche d'information, indexation sémantique, classification de documents biomédicaux, maladie d'Alzheimer

Title : Contribution to ontology building and to semantic information retrieval: application to medical domain

Abstract

This work aims at providing efficient access to relevant information among the increasing volume of digital data. Towards this end, we studied the benefit from using ontology to support an information retrieval (IR) system.
We first described a methodology for constructing ontologies. Thus, we proposed a mixed method which combines natural language processing techniques for extracting knowledge from text and the reuse of existing semantic resources for the conceptualization step. We have also developed a method for aligning terms in English and French in order to enrich terminologically the resulting ontology. The application of our methodology resulted in a bilingual ontology dedicated to Alzheimer's disease.
We then proposed algorithms for supporting ontology-based semantic IR. Thus, we used concepts from ontology for describing documents automatically and for query reformulation. We were particularly interested in: 1) the extraction of concepts from texts, 2) the disambiguation of terms, 3) the vectorial weighting schema adapted to concepts and 4) query expansion. These algorithms have been used to implement a semantic portal about Alzheimer's disease. Further, because the content of documents are not always fully available, we exploited incomplete information for identifying the concepts, which are relevant for indexing the whole content of documents. Toward this end, we have proposed two classification methods: the first is based on the k nearest neighbors' algorithm and the second on the explicit semantic analysis. The two methods have been evaluated on large standard collections of biomedical documents within an international challenge.

Keywords : ontology construction, TOR reuse, information retrieval, semantic indexing, biomedical document classification, Alzheimer's disease

Unité de recherche : ERIAS/Inserm U897

Remerciements

Je tiens à remercier Monsieur Roger Salamon, Directeur de l'équipe ERIAS, d'avoir accepté de diriger ce travail.

Je remercie spécialement mes co-encadrants Monsieur Gayo Diallo et Madame Fleur Mougin, pour leur grande disponibilité, leurs conseils et critiques constructives. Ils ont beaucoup contribué à l'aboutissement de cette thèse.

Je voudrais remercier Messieurs Moussa Lo, Professeur à l'Université Gaston Berger de Saint Louis et Pierre Zweigenbaum, Directeur de Recherche au CNRS, d'avoir accepté d'être les rapporteurs de ma thèse. Merci pour vos pertinentes remarques et suggestions.

Je remercie également Monsieur Jean François Dartigues et Guy Melançon, Professeurs à l'Université de Bordeaux et Madame Nathalie Hernandez, Maître de Conférences à l'Université de Toulouse, d'avoir accepté de participer au jury de ma thèse.

Je remercie aussi la fondation Plan Alzheimer d'avoir financé et soutenu ce travail.

Mes remerciements vont aussi à l'endroit des membres de l'équipe Epidémiologie et Neuropsychologie du Vieillissement Cérébral, dirigé par Jean François Dartigues, et à Madame Evelyne Mouillet pour leur collaboration.

Je tiens à remercier particulièrement mes parents pour leur soutien.

Mention spéciale à ma femme, ma famille et mes amis pour leur soutien et leurs encouragements.

Merci également aux membres de l'équipe ERIAS et à mes collègues doctorants de l'ISPED.

Je remercie enfin tous ceux qui, de près ou de loin, ont contribué à la réussite de ce travail.

Table des Matières

Liste des figures

Liste des tableaux

Introduction

1 Contextes

1.1 Contexte général

Le principe d'un système de recherche d'information (SRI) est de fournir, à partir d'une collection de documents, ceux qui sont pertinents pour une requête d'un utilisateur. Actuellement, une énorme quantité d'informations est disponible dans toute sorte de domaines. Ce volume d'informations croissant avec une production abondante de données numériques (articles scientifiques, portails d'information, comptes rendus, etc.) influe sur les performances des SRI. De plus, ces informations sont généralement exprimées en langage naturel (sous forme de textes) et donc dans un format non structuré (près de 80% de données dans le domaine médical selon les estimations) ; ce qui rend leur traitement automatique difficile. En outre, les utilisateurs peuvent avoir des niveaux d'expertise et des profils différents, allant des spécialistes du domaine traité à des utilisateurs grand public. D'autre part, l'accès à la bonne information est primordial pour aider à la prise de décision éclairée. Face à cet enjeu, il est nécessaire de concevoir des outils de recherche d'information (RI) plus robustes permettant aux utilisateurs de trouver facilement les informations pertinentes correspondant à leurs besoins.

Ces différentes questions ont motivé les travaux de la communauté RI qui ont abouti au développement de nombreux modèles et outils, mais également de méthodologies avancées pour leur évaluation.

Toutefois, les approches classiques de RI sont basées sur des mots clés; les documents et les requêtes sont décrits par un ensemble de mots (ou groupe de mots ou n-grammes; nous parlerons de termes dans la suite du document) qu'ils contiennent. Ensuite, la correspondance entre un document et une requête est basée sur le nombre de termes qu'ils partagent. Pour qu'un document soit sélectionné, il doit contenir les mêmes termes (ou une partie) que la requête de l'utilisateur. Ainsi, plus le nombre de termes en commun entre le document et la requête est élevé, plus sa pertinence par rapport à cette dernière est importante. Or, les documents pertinents ne contiennent pas toujours les mêmes termes que la requête (problématique de la synonymie). De même, les documents contenant les mêmes termes que la requête ne sont pas forcément pertinents (problématiques d'homonymie et d'ambiguïté). Ces approches restent ainsi confrontées à deux principaux problèmes linguistiques : la disparité et l'ambiguïté des termes. Enfin, la RI classique considère généralement les termes d'indexation comme des entités indépendantes et par conséquent ne permet pas la prise en compte des relations entre ces termes.

La recherche sémantique, définie comme la recherche basée sur la sémantique des termes, a été proposée pour surmonter ces problèmes et améliorer ainsi les performances des SRI traditionnels que nous qualifierons de « classiques ». L'idée est de prendre en compte le contenu sémantique véhiculé par les documents et les requêtes plutôt que de les décrire par de simples « sacs de mots ». Vu l'intérêt qu'elle a suscité, la RI sémantique a fait l'objet de

nombreux travaux de recherche ces dernières années. L'indexation conceptuelle utilisée pour décrire explicitement les documents par des concepts a connu un réel engouement. On peut distinguer deux catégories : 1) les approches statistiques qui exploitent la cooccurrence des termes dans un corpus pour définir des concepts en utilisant une technique de réduction de dimensions (Deerwester et al., 1990a) et 2) les approches basées sur des ressources sémantiques, telles que les thésaurus (Gonzalo et al., 1998) ou les ontologies (Kiryakov et al., 2004; Castells et al., 2007), dans lesquelles les concepts sont définis explicitement. Notre travail s'inscrit dans la deuxième approche et propose d'explorer le potentiel des ontologies pour guider un modèle de RI. En effet, l'ontologie constitue un modèle conceptuel idéal pour représenter l'information (les documents) mais aussi pour exprimer les besoins en information (les requêtes). Elle permet également d'exploiter les relations sémantiques structurant ces concepts pour améliorer les performances d'un SRI. La reformulation et l'expansion des requêtes initiales de l'utilisateur par des concepts sémantiquement proches permettent aussi d'améliorer la pertinence des réponses fournies par le système. Par exemple, pour un utilisateur qui désire accéder à des documents traitant de la « démence », le SRI peut lui proposer des documents traitant également de la « maladie d'Alzheimer », en exploitant les liens de spécialisation. Par ailleurs, l'ontologie permet d'unifier des termes synonymes dans des langues différentes au sein d'une même notion (ou concept). Ceci est ainsi particulièrement adapté et essentiel dans un contexte de RI multilingue.

1.2 Contexte d'application de notre travail

Le travail que nous avons mené dans le cadre de cette thèse trouve son contexte d'application privilégié dans le domaine de la maladie d'Alzheimer.

La mesure 32 du *plan Alzheimer* vise à améliorer la formation des médecins en épidémiologie clinique. Le diagnostic de la maladie d'Alzheimer étant complexe et son évolution rapide, une formation en épidémiologie clinique devient nécessaire pour le développement de la recherche. Le bulletin bibliographique BiblioDémences[1] initié par l'équipe « Epidémiologie et Neuropsychologie du Vieillissement Cérébral » du centre INSERM U897, s'inscrit dans ce cadre et offre des analyses critiques d'articles scientifiques portant sur la maladie d'Alzheimer et les syndromes apparentés. À partir d'une veille bibliographique hebdomadaire, une trentaine d'articles sont sélectionnés mensuellement à partir de la littérature scientifique mondiale et proposés à des spécialistes pour qu'ils en fassent une analyse critique. Les résumés (en anglais, principalement) de ces articles sont intégrés dans la base bibliographique BiblioDem[2] avec la lecture critique associée (en français) et quelques mots clés (environ quatre ou cinq) attribués manuellement par une documentaliste; les articles extraits de MEDLINE sont aussi associés à des descripteurs MeSH pour leur indexation. BiblioDem est une base cumulative qui contient actuellement plus de 1600 documents analysés et est augmentée chaque mois de nouveaux articles présentés et discutés lors d'un comité de lecture. Sa fonctionnalité de recherche utilise les mots clés et n'exploite qu'une partie des informations de la base. Elle ne permet pas d'assister les utilisateurs néophytes, pour leur faciliter l'accès aux informations pertinentes. Elle est donc insuffisante pour l'exploitation

[1] http://www.isped.u-bordeaux2.fr/CDD/FR_HTML_BIBLIONET.aspx
[2] http://www.isped.u-bordeaux2.fr/CDD/FR_HTML_BIBLIONET.aspx#BiblioDem

optimale d'une base bibliographique aussi riche que BibliDem. L'accès aux connaissances sur la maladie reste donc un enjeu important pour ses divers utilisateurs (chercheurs, étudiants en médecine, professionnels de santé, etc.). D'où le besoin d'un outil de recherche plus performant permettant aux utilisateurs de trouver facilement et rapidement les documents pertinents correspondant à leurs besoins.

Dans ce contexte, la modélisation et la formalisation des connaissances disponibles sur la maladie d'Alzheimer sont préalablement cruciales. Ensuite, un modèle de RI guidé par ce modèle de connaissances résultant permettrait une exploitation efficace de ces informations.

2 Problématique

L'intérêt suscité par la RI sémantique, en plus de l'abondance des ressources termino-ontologiques (RTO) dans le domaine biomédical, a motivé leur utilisation de plus en plus pour supporter des modèles de RI. Ainsi, des ressources sémantiques, telles que le thésaurus MeSH (Medical Subject Heading), ont été largement utilisées pour indexer des textes biomédicaux (Trieschnigg et al., 2009) et pour améliorer les performances de la RI par des techniques d'expansion de requêtes (Díaz-Galiano et al., 2009; Azcárate et al., 2012).

Toutefois, bien que de nombreuses ressources sémantiques aient été développées dans le domaine biomédical, il y a encore un besoin pour couvrir des domaines spécifiques comme la maladie d'Alzheimer où les connaissances du domaine sont régulièrement enrichies. L'accès aux connaissances sur cette maladie reste aujourd'hui un enjeu important pour les chercheurs, les praticiens mais également les décideurs politiques, nécessitant ainsi de modéliser les connaissances du domaine. Une telle modélisation peut se faire via une ontologie de domaine (Guarino et Giaretta, 1995). Par ailleurs, pour que ce modèle de connaissances permette à la communauté scientifique internationale, et plus particulièrement aux experts du domaine, de partager facilement et largement leurs connaissances sur la maladie et les syndromes associés, l'idéal serait qu'il soit multilingue. Notons cependant que la construction d'ontologies de domaine reste une tâche fastidieuse et coûteuse, qui nécessite une méthodologie élaborée permettant d'alléger le processus. Aucune approche standard n'existe comme c'est le cas dans le domaine des bases de données.

D'autre part, même si d'importants travaux ont été réalisés dans la RI sémantique biomédicale, de nombreuses questions restent ouvertes. L'utilisation des ressources sémantiques en RI soulève, en effet, de nombreux challenges que sont : la disponibilité de ressources adaptées pour le domaine d'application, l'identification des descripteurs importants (concepts) dans des textes (Suominen et al., 2013) pour des tâches d'indexation, la sélection automatique de ceux qui sont les plus pertinents pour représenter les documents (on parle aussi de *classification*), la couverture suffisante des ressources (Bhagdev et al., 2008), etc. De plus, bien que beaucoup de travaux récents aient contribué à développer plus largement et à améliorer l'indexation sémantique de documents biomédicaux, elle reste encore un véritable défi. En particulier, les textes intégraux des documents n'étant pas toujours accessibles, déterminer les concepts pertinents permettant de décrire le contenu des documents complets en se basant seulement sur une partie des informations disponibles, telles que les titres et les résumés, est encore un challenge (Tsatsaronis et al., 2012).

L'objectif de cette thèse est double. Il s'agit, d'une part, de réutiliser ou développer une approche de construction d'ontologies afin de mettre en œuvre une ontologie du domaine et, d'autre part, de proposer une approche de RI sémantique basée sur cette ontologie. Notre cas d'application est la maladie d'Alzheimer et ses syndromes apparentés. Les documents de la base BiblioDem constituent la source d'information permettant de concevoir l'ontologie de domaine, que nous appellerons OntoAD, et le modèle de RI s'attachera à réaliser l'indexation conceptuelle de ces documents pour les rendre disponibles via un portail sémantique de RI, que nous appellerons SemBiP (Semantic BiblioDem Portal).

Ce travail soulève donc les questionnements scientifiques sous-jacents suivants:

- Quelle approche méthodologique de construction d'ontologies est-elle appropriée pour notre problème ?
- Comment tirer profit des RTO existantes susceptibles d'être réutilisées pour alléger le processus de construction d'ontologies ?
- Comment exploiter le langage des spécialistes (résumés et lectures critiques dans différentes langues) pour la conceptualisation et la formalisation de leurs connaissances sous forme d'une ontologie ?
- Comment gérer l'aspect multilingue de l'ontologie ? Quelle stratégie pour aligner des termes synonymes dans des langues différentes ?
- Une fois construite et compte tenu du caractère évolutif du domaine d'application, comment gérer la maintenance et l'évolution de l'ontologie et quel mécanisme de persistance choisir pour en faciliter la gestion ?
- Comment traiter les défis soulevés par la RI basée sur une ontologie que sont notamment le repérage des concepts, leur éventuelle désambiguïsation, la couverture potentiellement limitée de l'ontologie, la sélection des concepts pertinents pour indexer un document ?
- Comment indexer un document dont le contenu n'est disponible que partiellement (par exemple, uniquement les résumés) ?
- Comment utiliser une telle ontologie pour supporter la gestion d'un portail sémantique?
- Comment prendre en compte différentes modalités d'accès au portail sémantique (navigation, recherche avec les concepts, recherche en texte libre, etc.) ?
- Comment évaluer l'apport de l'ontologie au sein du portail sémantique ? Comment évaluer ce portail à part entière ?

Le travail mené dans le cadre de cette thèse a pour but de contribuer à répondre à ces questionnements.

3 Plan du mémoire

Ce mémoire comprend deux parties. La première partie s'intéresse à l'état de l'art et est subdivisée en deux chapitres : le premier présente un état de l'art sur la construction d'ontologie tandis que le deuxième expose les travaux existants en RI sémantique. La deuxième partie présente nos différentes contributions. Elle comprend trois chapitres

décrivant notre approche de construction d'ontologie, notre méthode d'indexation conceptuelle et de RI et les approches que nous proposons pour la classification d'une large collection de documents biomédicaux dont le contenu exhaustif n'est pas disponible.

Dans le **chapitre 1**, après un aperçu des différentes notions associées aux « ontologies », nous décrivons les différents types de modèles de connaissances mais aussi les langages permettant de les représenter et de les manipuler. Nous présentons ensuite certaines RTO développées dans le domaine médical. Enfin, nous faisons une revue sur les travaux concernant la construction d'ontologies. Nous nous intéressons plus particulièrement aux méthodologies d'acquisition d'ontologies à partir de textes et celles basées sur la réutilisation de RTO existantes.

Dans le **chapitre 2**, nous passons en revue les travaux sur la RI en nous focalisant sur les approches sémantiques. Ainsi, après une description du fonctionnement des SRI, les modèles de RI les plus courants sont présentés. Ensuite, des méthodes et campagnes proposées dans la littérature pour leur évaluation sont décrites. Nous terminons par une présentation des travaux sur la RI sémantique, particulièrement ceux basés sur des ressources sémantiques.

Le **chapitre 3** est consacré à la description de notre approche de construction d'ontologies qui combine des techniques de traitement automatique des langues (TAL) et la réutilisation de RTO existantes. Les différents outils et ressources sont d'abord présentés. Ensuite, nous détaillons les différentes étapes et l'illustrons par la mise en œuvre d'une ontologie bilingue de la maladie d'Alzheimer.

Le **chapitre 4** s'intéresse à différentes questions soulevées par les approches de RI basées sur des ressources sémantiques : l'identification de concepts, la désambiguïsation, l'expansion des requêtes et la couverture limitée des ressources. Nous proposons ainsi deux stratégies d'extraction de concepts à partir de textes. Une méthode de désambiguïsation exploitant la similarité sémantique est aussi décrite. Pour faire face à l'incomplétude des ressources sémantiques, nous proposons la combinaison de la RI sémantique et la recherche par mots clés. Nous présentons également l'application de nos différentes propositions pour la mise en œuvre d'un portail sémantique dédié à la maladie d'Alzheimer, SemBiP.

Dans le **chapitre 5**, nous exposons deux méthodes qui, à partir des informations partielles disponibles pour des documents (titre, résumé), permettent de déterminer les concepts pertinents permettant de les représenter. La première utilise l'approche des k plus proches voisins tandis que la seconde se base sur l'analyse sémantique explicite. Une évaluation de ces deux méthodes sur des collections standards et une analyse des différents résultats sont aussi décrites.

Dans le **chapitre 6**, nous discutons les limites de nos propositions et les perspectives de ce travail.

Chapitre 1: Etat de l'art sur les approches de construction d'ontologies

1 Introduction

Les ontologies ont aujourd'hui une place majeure dans la représentation et la modélisation des connaissances. Elles sont utilisées pour formaliser les connaissances d'un domaine et ainsi ajouter une couche sémantique aux systèmes et applications informatiques. Les ontologies permettent de représenter de manière explicite les connaissances d'un domaine au moyen d'un langage formel afin qu'elles puissent être manipulées automatiquement et partagées aisément. Leur utilisation est largement répandue et ce, dans divers domaines de recherche tels que la représentation de connaissances (Hoehndorf et al., 2014), la RI (Castells et al., 2007) et l'intégration de données (Wache et al., 2001).

Dans ce chapitre, nous allons présenter les différentes notions liées aux ontologies. Nous exposons ensuite des travaux sur les différentes manières de construire de telles ressources.

Ainsi, nous commençons d'abord par donner quelques définitions courantes d'ontologies en ingénierie des connaissances (section 2) avant de décrire leurs différents composants (section 3) puis de présenter leur typologie (section 4). Les langages de représentation et de manipulation d'ontologies sont ensuite décrits dans la section 5. Une brève description des ressources termino-ontologiques (RTO) disponibles dans le domaine médical est faite dans la section 6. La section 7 revisite les différentes approches de construction d'ontologies en se focalisant sur celles qui exploitent des ressources existantes. Nous terminons par une conclusion dans la section 8.

2 Définitions

La notion d'ontologie est utilisée dans des contextes différents touchant la philosophie, où le terme a été initialement introduit, et l'intelligence artificielle, l'ingénierie des connaissances (IC) en particulier.

Dans le premier usage, l'ontologie est la partie de la philosophie qui s'intéresse à la nature et à l'organisation de la réalité (Guarino et Giaretta, 1995). Elle est la science de l'être dans laquelle on cherche à représenter le monde.

Dans le domaine IC, la notion d'ontologie a été introduite dans les années 1990 avec les travaux de Tom Grüber (Grüber, 1993a) qui a proposé la définition, certainement la plus populaire, parmi celles qui existent dans la littérature : *une ontologie est une spécification explicite d'une conceptualisation*. Une *conceptualisation* est une vue abstraite des entités réelles du monde que l'on veut représenter et de leurs relations; c'est un modèle abstrait des phénomènes réels du monde. L'expression *spécification explicite* signifie que la conceptualisation est faite de façon non ambiguë dans un langage concret; les concepts et les contraintes utilisés pour représenter le domaine de connaissances doivent être explicitement définis. La définition de Gruber a ensuite été complétée par Studer et ses collègues (Studer et

al., 1998). Ils définissent une ontologie comme une *spécification formelle et explicite d'une conceptualisation partagée*. Le terme *formelle* fait référence au fait qu'une ontologie doit être compréhensible par les machines, c'est-à-dire que ces dernières doivent être capables d'interpréter la sémantique de l'information fournie. Une ontologie doit aussi représenter une connaissance consensuelle *partagée* et acceptée par une communauté d'utilisateurs afin d'en garantir une utilisation plus large.

Cette définition est similaire à celle donnée par Guarino (Guarino, 1998) qui considère une ontologie comme un vocabulaire spécifique utilisé pour décrire une certaine réalité, ainsi qu'une spécification sur le sens des termes de ce vocabulaire. Ainsi, une ontologie permet une description formelle des connaissances d'un domaine, réel ou imaginaire, dans le but de les rendre permanentes, de faciliter leur partage et leur exploitation automatique par des machines.

Selon Uschold, *une ontologie peut prendre différentes formes, mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. Cette dernière inclut des définitions et une indication de la façon dont les concepts sont reliés entre eux, les liens imposant collectivement une structure sur le domaine et contraignant les interprétations possibles des termes* (Uschold, 1998). Une ontologie comprend donc un vocabulaire commun aux utilisateurs matérialisé par des **concepts** ainsi que des **relations** entre ces concepts. Autrement dit, une ontologie est vue comme un ensemble de concepts et de propriétés caractérisant les objets d'un domaine du monde réel. Ces concepts sont organisés sous la forme d'un graphe et structurés par des relations hiérarchiques (ou taxonomiques) et associatives (ou transversales). En plus des concepts et relations, elle définit des **règles** ou **contraintes** sur leurs définitions et interprétations permettant d'effectuer des déductions.

Roche a résumé ces différentes définitions : *définie pour un objectif donné et un domaine particulier, une ontologie est pour l'ingénierie des connaissances une représentation d'une modélisation d'un domaine partagée par une communauté d'acteurs. Objet informatique défini à l'aide d'un formalisme de représentation, elle se compose principalement d'un ensemble de concepts définis en compréhension, de relations et de propriétés logiqu*es. (Roche, 2005).

Nous retiendrons donc qu'une ontologie est constituée d'un ensemble de concepts et de relations entre ces concepts. Elle peut également comprendre des contraintes.

Les ontologies sont utilisées dans différentes applications et notamment dans les systèmes à bases de connaissances (BC) et le Web sémantique (Berners-Lee et al., 2001). Grace à leur capacité de description formelle des connaissances d'un ou plusieurs domaines, les ontologies constituent un outil intéressant pour supporter les systèmes à base de connaissances. Puisque la mise en place d'une BC n'est pas simple, les ontologies semblent appropriées pour faciliter leur construction, leur réutilisation et leur partage (Mizoguchi et al., 1995). D'autre part, le Web sémantique dont l'objectif est de développer des standards et technologies pour permettre aux machines de « comprendre » et de traiter de grands volumes d'informations et de services disponibles sur le Web (Berners-Lee et al., 2001) est une application où les

ontologies sont largement utilisées. En effet, la plupart des informations disponibles sur le Web sont compréhensibles par des humains mais elles étaient jusque-là difficilement interprétables par les machines, ce qui pose problème pour la recherche et le partage d'informations. Afin de permettre à des agents et programmes informatiques de pouvoir interpréter ces informations et de communiquer avec d'autres programmes pour leur exploitation de manière intelligente, ces dernières nécessitent d'être décrites explicitement. Dans ce cadre, les ontologies sont un système conceptuel approprié pour décrire les ressources. Elles sont donc un élément crucial et essentiel du Web sémantique. Par conséquent, l'émergence du Web sémantique a beaucoup motivé le développement d'ontologies.

En fonction de l'usage qui en est fait, une ontologie peut être constituée de différents composants. Dans la section suivante, nous décrivons les principaux éléments constitutifs d'une ontologie.

3 Les composants d'une ontologie

D'après les définitions précédentes, une ontologie est composée principalement des éléments suivants : des concepts, des relations entre ces concepts, des instances de concepts et des axiomes (Gómez-Pérez, 1999; Noy et Mcguinness, 2001). Pour illustrer les différents composants d'une ontologie, nous introduirons des exemples simples du domaine de l'enseignement et la recherche scientifique.

3.1 Concepts et instances

Un concept (ou classe au sens des standards du Web Sémantique que nous présenterons rapidement par la suite) représente un ensemble d'objets et leurs propriétés communes ; il peut représenter un objet matériel, un événement, une notion ou une idée (Uschold et King, 1995). Un concept est défini par un terme et a généralement une **intension** et/ou une **extension**. L'intension désigne la sémantique, c'est-à-dire l'ensemble des attributs et propriétés définissant un concept. L'ensemble des **objets** qu'englobe un concept est son extension. Des exemples de concepts sont : *Personne*, *Etudiant*, *Enseignant*, *Chercheur*, *Université* et *Cours* où le concept *Chercheur* peut être défini suivant :

- son intension : une personne dont le métier consiste à faire de la recherche scientifique;
- son extension : l'ensemble des personnes qui obéissent à cette définition.

Ces objets, décrits par un concept, sont appelés ***instances*** du concept. Certains travaux considèrent les instances comme des composants d'une ontologie tandis que d'autres considèrent la somme d'une ontologie et des instances de ses concepts comme une base de connaissances (Noy et Mcguinness, 2001). Des exemples d'instances sont : Gayo Diallo (instance de *Chercheur*), le cours de Web sémantique de Gayo Diallo (instance de *Cours*), l'université de Bordeaux (instance d'*Université*).

Les concepts peuvent également disposer d'attributs permettant de décrire leurs caractéristiques. Par exemple, le concept *Etudiant* dispose des propriétés *numéro étudiant, nom, prénom, date de naissance* et *adresse*.

Enfin, il peut exister des propriétés telles que l'***équivalence*** pour exprimer que deux concepts représentent la même chose ou la ***disjonction*** pour exprimer l'incompatibilité entre deux concepts. Par exemple, les concepts *Enseignant* et *Cours* sont disjoints.

3.2 Relations

Au sein d'une ontologie, les concepts ainsi que les instances peuvent être reliés entre eux par des relations. Une relation est un lien (binaire, tertiaire, etc.) entre des entités, exprimée souvent par un terme ou un symbole littéral. Elle peut être caractérisée par une signature qui précise le nombre d'instances de concepts que la relation lie, leurs types et l'ordre des concepts. Par exemple, la relation *enseigner* lie une instance du concept *Enseignant* à au moins une instance du concept *Cours* (dans cet ordre). Dans cet exemple, *Enseignant* est appelé le **domaine** et *Cours* le **co-domaine** de la relation *enseigner*.

Une ontologie est souvent représentée par une hiérarchie de concepts où ces derniers sont structurés suivant des relations de subsomption (relations « *est un* » ou « *est une sorte de* »). Cette relation, dite taxonomique, lie un élément supérieur (le concept plus général) et un élément inférieur (le concept spécifique). Un concept C_1 subsume un concept C_2 si l'extension de C_2 est incluse dans l'extension de C_1. Autrement dit, toute propriété sémantique de C_1 est aussi une propriété sémantique de C_2, ou encore C_2 est plus spécifique que C_1. Par exemple, le concept *Personne* subsume le concept *Enseignan*t que l'on peut également exprimer en disant que le concept *Enseignant* spécialise le concept *Personne*.

Les relations de méronymie (de type « *partie de* ») sont également utilisées pour structurer les concepts d'une ontologie. C'est une relation qui lie un couple de concepts dont l'un est une partie de l'autre. Dans une relation de méronymie, les propriétés du tout ne sont pas obligatoirement transmises à ses parties. Par exemple, un *Enseignant* ou un *Etudiant* font partie d'une *Université* sans que les propriétés d'*Université* soient transmises à *Enseignant* ou *Etudiant*.

En plus des relations hiérarchiques et partitives, les concepts dans une ontologie peuvent entretenir d'autres types de relations : on parle de relations « associatives » ou « transversales ». Par exemple, une relation *suivre* peut être définie entre le concept *Etudiant* et le concept *Cours*.

3.3 Axiomes

Ce sont des contraintes ou des règles sur des concepts (on parle dans ce cas d'axiome) ou des instances (on parle dans cet autre cas d'assertion) de l'ontologie considérées comme toujours vraies. Les axiomes permettent d'inférer de nouvelles connaissances à partir d'un ensemble de faits. Par exemple, on peut définir une règle du type : un *Enseignant* qui *donne* un *Cours* sur un *Thème* « *connait* » ce *Thème*. Si l'on décrit ensuite un enseignant *E* qui donne un cours *C* portant sur le « *Web sémantique* », alors il sera possible de déduire que *E* connait le « *Web sémantique* ». Contrairement aux concepts et relations, l'expression des axiomes nécessite un langage qui supporte le raisonnement.

4 Typologie des ontologies

Différentes typologies ont été proposées pour distinguer les multiples ontologies existantes : selon leur couverture, en fonction de leur granularité ou encore de leur niveau de formalisation. Parallèlement, l'engouement généré par les multiples fonctionnalités offertes par les ontologies a résulté en un développement très rapide de ressources « dérivées » ne respectant pas exactement la définition et le mode d'utilisation des ontologies, en particulier dans le domaine médical (Soldatova et King, 2005).

4.1 Typologie selon l'objet de la conceptualisation

En fonction de leur portée et des objectifs visés, différents types d'ontologies sont développés (Diallo, 2006; Mizoguchi, 2003) : les ontologies de représentation, les ontologies de haut niveau, les ontologies génériques, les ontologies de domaine et les ontologies d'application.

Les **ontologies de représentation** définissent les concepts impliqués dans la formalisation des connaissances. Elles fournissent les primitives nécessaires pour décrire les concepts des autres types d'ontologies. Un exemple d'ontologie de ce type est la Frame Ontology qui rassemble les primitives de représentation des langages à base de frame : *classe*s, *instances*, *propriétés*, *relations*, *restrictions*, etc. (Gruber, 1993b). Les ontologies de représentation permettent de définir les langages de représentation des connaissances.

Les **ontologies de haut niveau** ou *upper-level ontologie*s (ontologies de haut niveau) visent à modéliser les concepts de haut niveau d'abstraction (*entity*, *event*, *process*, etc). Les ontologies Basic Formal Ontology (BFO) (Grenon et Smith, 2004), Suggested Upper Merged Ontology (SUMO) (Niles et Pease, 2001) et Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Gangemi et al., 2002) sont des exemples de ce type d'ontologies.

Les **ontologies génériques**, appelées également *méta-ontologies* ou *core-ontologies*, décrivent des concepts généraux, indépendants d'un domaine ou d'un problème particulier. Elles décrivent des concepts génériques valables dans différents domaines mais moins abstraits que ceux décrits dans les ontologies de haut niveau. L'ontologie de Sowa (Sowa, 2000) et CYC (Lenat et Guha, 1989) sont des exemples d'ontologies génériques.

Les **ontologies de domaine** décrivent des connaissances d'un domaine spécifique. Leurs concepts sont souvent définis comme une spécialisation des concepts des ontologies de haut niveau. Les ontologies de domaine sont définies de deux manières :

- des ontologies spécifiques à un domaine particulier qui modélisent les connaissances spécifiques à ce domaine précis. Des exemples d'ontologie de ce type sont le Foundational Model of Anatomy (FMA) qui décrit la structure macroscopique du corps humain (Rosse et Mejino Jr., 2003) et l'ontologie IDOSCHISTO qui représente des connaissances sur la bilharziose (Camara et al., 2013);
- des ontologies de tâches ou d'application qui concernent des tâches réalisées dans un domaine particulier. Elles décrivent un vocabulaire en relation avec une tâche ou une activité d'un domaine. Généralement, ces ontologies ne sont pas réutilisables et

possèdent une portée limitée. On peut citer par exemple l'ontologie « The scheduling task ontology » (Rajpathak et al., 2001) qui vise à construire des applications pour gérer les emplois du temps.

Les ontologies de domaine permettent la représentation des connaissances d'un domaine particulier et leur réutilisation par des applications de ce domaine. C'est le type d'ontologies le plus courant en ingénierie ontologique.

4.2 Typologie des ontologies selon leur granularité

Selon le niveau de granularité, on peut distinguer deux types d'ontologies (Diallo, 2006):

- **Granularité fine** : c'est lorsque l'ontologie est très détaillée. Elle permet de décrire précisément les connaissances. Souvent les ontologies de domaine, les ontologies de tâches et les ontologies d'application sont de ce type.
- **Granularité large** : dans le cas où l'ontologie est moins détaillée, on parle de granularité large (par exemple, les ontologies de haut niveau). Ce type d'ontologie peut être partagée et raffinée dans d'autres types d'ontologies telles que les ontologies de domaine ou d'application.

4.3 Typologie des ontologies selon leur niveau de formalisation

On peut également classifier les ontologies en fonction de leur niveau de formalisation. Ainsi, on peut distinguer :

- Les **ontologies formelles** basées sur un langage artificiel possédant une sémantique formelle ;
- Les **ontologies informelles** exprimées en langage naturel.

Studer et ses collègues distinguent aussi les ontologies légères (*light-weight ontologies*) des ontologies lourdes (*heavy-weight ontologies*) (Studer et al., 1998).

- Une **ontologie légère** comprend un ensemble de concepts structurés par des relations hiérarchiques et associatives. C'est le type d'ontologie le plus couramment utilisé, notamment en RI (Chrisment et al., 2008).
- Une **ontologie lourde** comprend en plus des axiomes qui permettent de définir des expressions complexes (contraintes sur les concepts et/ou relations) et suppose l'existence d'un système de déduction. Ce type d'ontologie est plus expressif mais sa mise en place est souvent coûteuse.

4.4 Les autres ressources de connaissances

A côté de cette classification des ontologies, il y a des ressources terminologiques (Aussenac-Gilles et al., 2002; Roche, 2012; de Keizer et al., 2000) sémantiquement proches et parfois considérées, en fonction des applications et des domaines, comme des variantes ontologiques : le vocabulaire contrôlé, la terminologie, la taxonomie et le thésaurus.

Un **vocabulaire contrôlé** : c'est un ensemble de termes reconnus, normalisés et validés par une communauté de pratiques, le vocabulaire technique d'un domaine de spécialité. Les termes doivent être définis de manière non ambiguë. Si plusieurs termes désignent la même

notion (« concept »), l'un d'entre eux représente le terme préféré et les autres ses synonymes. Les vocabulaires contrôlés sont souvent utilisés dans l'indexation contrôlée et la RI.

Une **terminologie** : c'est un vocabulaire contrôlé dont les termes sont associés à des significations. Elle peut être considéré comme une langue de spécialité pour définir les termes d'un domaine particulier (Roche, 2005). Les termes sont liés par la relation d'hyperonymie. *La terminologie ne s'intéresse aux mots que dans la mesure où ils désignent des notions (ou concepts)*… C'est un *système de termes reflétant une modélisation conceptuelle*. Donc, similairement à l'ontologie, son but est de permettre la compréhension du monde, et de faciliter ainsi la communication et le partage d'informations. Ainsi, les définitions des termes dans une terminologie doivent être consensuelles et acceptées par une communauté ; elles doivent être claires et non ambiguës.

Une **taxonomie** de concepts : c'est un ensemble de concepts structurés par des relations de subsomption, i.e. chaque instance d'une classe est aussi instance de ses superclasses.

Un **thésaurus** : c'est un réseau de termes organisé avec plusieurs types de relations : des relations hiérarchiques (généralisation/spécification), d'équivalence (synonymie) et des relations d'association entre termes. La hiérarchie est basée sur les relations « *est plus général* » (Broader than) et « *est plus spécifique* » (Narrower than) et ne correspond donc pas à une subsomption. Chaque « concept » est représenté par un terme préféré (Preferred Term) et des termes synonymes ou quasi-synonymes. Les termes sont également liés par des relations non hiérarchiques (Related Term).

Dans la suite, nous utilisons de manière générique le terme d'ontologie ou de *ressources termino-ontologiques* (RTO) (Charlet et al., 2012) pour désigner ces différents types de ressources, particulièrement répandues dans le domaine médical.

Dans la prochaine section, nous présentons les langages de représentation des connaissances, qui sont essentiels pour décrire une ontologie formelle.

5 Les langages de représentation et de manipulation d'ontologies

Plusieurs langages de représentation et de manipulation d'ontologies ont été développés. Dans cette section, nous faisons une rapide revue de ceux qui nous paraissent très représentatifs parmi les standards et recommandations du World Wide Web Consortium (W3C)[3] : RDF/RDFS, OWL, OWL 2, SKOS et SPARQL.

5.1 RDF (Resource Description Framework)

RDF[4] (Resource Description Framework) est un des formalismes standards de représentation de connaissances (Lassila et al., 1998). C'est une recommandation du W3C pour décrire des ressources et leurs **métadonnées** afin de faciliter leur traitement automatique. Basé sur la syntaxe XML (eXtended Markup Language), le langage RDF permet de décrire des

[3] http://www.w3.org/
[4] http://www.w3.org/TR/2002/WD-rdf-concepts-20021108/

connaissances sous forme de ressources, propriétés et valeurs. En d'autres termes, c'est un formalisme utilisé pour représenter les propriétés d'une ressource et leurs valeurs. Un document RDF est représenté par une collection de triplets de la forme <sujet, propriété, valeur> (ou <sujet, prédicat, objet>). Le sujet est toujours une ressource identifiée par son IRI (Internationalized Resource Identifier). Une propriété associe une valeur à une ressource (le sujet). Les propriétés concernent les attributs caractérisant les concepts et les relations entre concepts. Les valeurs peuvent être des ressources ou des littéraux. Par exemple, l'assertion l'enseignant E_1 enseigne le cours C_3 peut être représentée par le triplet <E_1, *enseigner*, C_3> où E_1 est le sujet, *enseigner* une propriété et C_3 la valeur de la propriété pour ce sujet.

Cet ensemble de triplets peut être représenté par un graphe dont les éléments apparaissant comme sujet ou objet sont les sommets, et où chaque triplet est représenté par un arc étiqueté entre deux sommets dont l'origine est son sujet et la destination sa valeur (voir Figure 1).

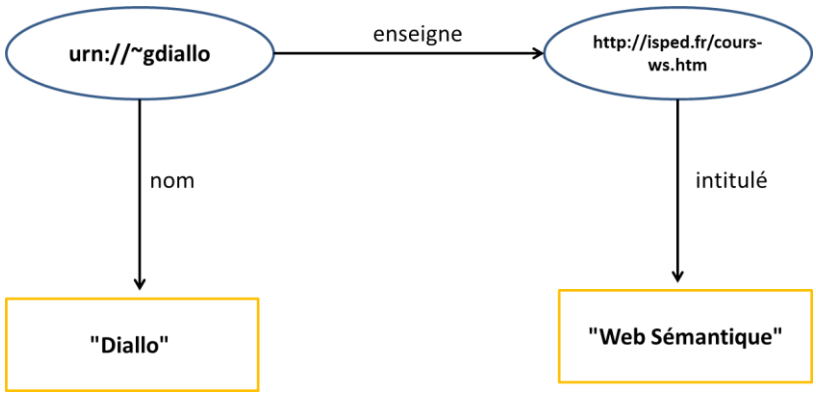


Figure 1 : Un exemple de modèle de graphe RDF

RDF étant avant tout un langage de description de métadonnées, il utilise le vocabulaire défini par RDFS pour cette description.

5.2 RDFS (Resource Description Framework Schema)

C'est un schéma de base incluant les entités sémantiques généralement utilisées (classes, sous-classes, propriétés, sous-propriétés, etc.) pour la structuration des connaissances d'un domaine . RDFS[5] est une extension de RDF permettant de définir les concepts utilisés dans les descriptions et les contraintes de type sur les objets et les valeurs des triplets. C'est un langage de définition de vocabulaires qui assure la description de classes et propriétés représentées dans des ressources RDF. En d'autres termes, les triplets RDF sont des instances de RDFS. Les classes et propriétés du domaine à modéliser sont définies respectivement comme des instances des primitives ***Class*** et ***Property*** de RDFS. Les primitives ***subClassOf*** et ***subPropertyOf*** permettent, quant à elles, de définir les relations hiérarchiques respectivement entre classes et entre propriétés. D'autres primitives comme ***domain*** et ***range*** sont aussi utilisées pour spécifier les domaines et co-domaines des relations. RDF et RDFS permettent ainsi la description de ressources sous forme de graphes de triplets. Cependant, ils souffrent de quelques limites puisqu'ils ne permettent pas d'exprimer :

[5] http://www.w3.org/TR/2002/WD-rdf-schema-20021112/

- des propriétés algébriques comme la **transitivité**, la **symétrie**. Par exemple, *aLeMêmeGradeQue* est symétrique, *aDeMeilleursRésultatsQue* est transitive.
- la combinaison booléenne de classes avec des opérateurs d'**union**, d'**intersection** et de **complémentarité**. Par exemple, *EnseignantChercheur* est l'union disjonctive de *ATER*, *MaîtreDeConférences* et *Professeur*.
- la **disjonction** entre classes : dire que deux classes sont disjointes. Par exemple, *PersonnelEnseignant* et *PersonnelAdministratif* sont disjoints.
- la restriction de **cardinalités** : c'est-à-dire le nombre de valeurs qu'une propriété donnée peut avoir. Par exemple, un *Enseignant enseigne au moins un Cours*.

Ainsi, ces deux langages ne permettent pas de représenter ce type d'axiomes et de les utiliser pour effectuer des déductions. En conséquence, un autre langage, OWL (Web Ontology Language), a été proposé pour combler ce manque.

5.3 OWL (Web Ontology Language)

Dans le but d'étendre l'expressivité du langage RDFS, le groupe de travail sur les ontologies du W3C, WebOnt[6], a recommandé le langage OWL[7] comme un standard pour la représentation d'ontologies depuis février 2004. OWL est un langage basé sur le langage DAML + OIL (McGuinness et al., 2002). En plus de ses constructeurs permettant de décrire des classes et des propriétés plus complexes, OWL intègre des outils de comparaison de propriétés et de classes. Il permet ainsi d'exprimer des propriétés telles que l'équivalence, la disjonction entre classes, la cardinalité, la symétrie, la transitivité d'une relation, etc. C'est un langage inspiré des logiques de description (Baader et al., 2003) qui offre aussi de nouvelles primitives permettant de définir des classes à l'aide de mécanismes ensemblistes (intersection de classes, union de classes, complément d'une classe). Il permet également d'effectuer des déductions. Ainsi, OWL peut être utilisé pour représenter explicitement une ontologie. De plus, en permettant la mise en relation d'ontologies différentes, OWL permet la prise en compte de l'aspect distribué des ontologies. Grâce à son pouvoir expressif et son pouvoir de raisonnement, il demeure un puissant langage de description d'ontologies.

OWL est doté de trois sous-langages de plus en plus expressifs:

- **OWL Lite** : c'est le sous langage le plus simple et le moins expressif. Il permet de représenter une hiérarchie de classes simples et d'exprimer des contraintes simples. Certaines propriétés comme la disjonction et l'union de classes ne sont pas supportées. Il supporte seulement une partie des constructeurs d'OWL.
- **OWL DL** (OWL Description Logic) : Il est plus complexe mais aussi plus expressif qu'OWL Lite. OWL DL supporte tous les constructeurs du langage OWL mais ils sont utilisés avec certaines restrictions. Par exemple, OWL DL ne permet pas la définition de relations transitives. Il se fonde sur la logique de description et supporte le raisonnement automatisé. OWL DL est caractérisé par la complétude de ses raisonnements (toutes les déductions peuvent être calculées) et leur décidabilité (calculs en un temps fini). Mais il n'est pas totalement compatible avec RDF/RDFS.

[6] http://www.w3.org/2001/sw/WebOnt/
[7] http://www.w3.org/TR/owl-features/

- **OWL Full** : C'est la version la plus complexe qui utilise toutes les constructions du langage OWL sans restrictions. C'est aussi le sous-langage qui assure le plus haut niveau d'expressivité. OWL Full est totalement compatible avec RDF. Il a l'avantage de pouvoir étendre le vocabulaire prédéfini (RDF ou OWL) mais son inconvénient majeur est la non-garantie de la complétude et de la décidabilité.

Entre ces trois sous-langages, il existe une compatibilité ascendante. Autrement dit, toute ontologie OWL Lite valide est également une ontologie OWL DL valide, et toute ontologie OWL DL valide est également une ontologie OWL Full valide. En fonction des besoins de l'utilisateur, un sous langage d'OWL peut être plus approprié qu'un autre.

5.4 OWL 2 (OWL 2 Web Ontology Language)

OWL 2[8] est une extension et une révision du langage OWL, développé par le groupe WebOnt du W3C. Il est devenu une recommandation du W3C depuis le 11 Décembre 2012. A l'instar d'OWL, ce langage a été conçu pour faciliter le développement d'ontologies et leur partage via le Web. Sa syntaxe principale est RDF/XML. Sa structure générale est très similaire à OWL. Ainsi, toutes les ontologies OWL restent des ontologies OWL 2 valides, avec des déductions identiques. Comparé à OWL, OWL 2 intègre de nouvelles fonctionnalités et l'étend par des simplifications syntaxiques et l'augmentation de son niveau d'expressivité. Par exemple, pour exprimer l'union disjointe : *DisjointUnion(:EnseignantChercheur :ATER :MaîtreDeConférences :Professeur).* Le langage OWL 2 comprend trois profils (sous-langages) différents : OWL 2 EL, OWL 2 QL, et OWL 2 RL. Chaque profil est défini comme une restriction syntaxique de la spécification structurelle d'OWL 2. Chaque profil a un niveau d'expressivité restreint qui lui permet de bénéficier de certains avantages en termes de performance ou de mise en œuvre :

- OWL 2 EL permet d'effectuer des tâches de raisonnement standards avec une complexité polynomiale du temps d'exécution. Ce sous-langage est adapté pour traiter les ontologies volumineuses.
- OWL 2 QL permet de traiter des requêtes conjonctives avec une complexité logarithmique. Ce profil est recommandé pour les applications utilisant des ontologies légères et où l'accès aux données se fait directement via des requêtes relationnelles.
- OWL 2 RL supporte les algorithmes de raisonnement avec un temps d'exécution polynomial. Il utilise les technologies d'extension de règles des bases de données opérant directement sur les triplets RDF. Ce sous-langage est particulièrement approprié pour des applications utilisant des ontologies relativement légères pour organiser un grand nombre d'individus et où il faut opérer directement sur les données sous la forme de triplets RDF.

5.5 SKOS (Simple Knowledge Organization System)

Recommandation du W3C depuis août 2009, SKOS[9] est un modèle de données destiné à supporter des RTO, telles que les terminologies, les thésaurus ou encore les taxonomies. De

[8] http://www.w3.org/TR/owl2-overview/
[9] http://www.w3.org/TR/skos-reference/

plus, il permet de décrire très en détail le niveau lexical d'une ressource ; ce qui est particulièrement intéressant dans des contextes de modélisation multilingue. Il offre un mécanisme simple permettant de supporter la représentation de vocabulaires structurés. Chaque concept est défini comme une classe RDFS avec un ensemble de propriétés; un concept possède un URI pour son identification, un terme préféré (ou un par langue dans le cas d'une ressource multilingue), et éventuellement des termes synonymes (ou termes alternatifs) et une (des) définition(s). Les termes (ou labels) sont des chaînes de caractères. Les concepts sont reliés entre eux suivant des relations de généralisation/spécialisation (*skos:broader* et *skos:narrower*) et associatives (*skos:related*). La figure 2 donne un exemple de représentation du concept *Chercheur* en SKOS.

```
onto:Chercheur rdf:type skos:Concept;
        skos:prefLabel "researcher"@en;
        skos:prefLabel "chercheur"@fr;
        skos:broader onto:Personne;
        skos:related onto:Thème;
        skos:related onto:Equipe.
onto:Personne rdf:type skos:Concept.
onto:Thème rdf:type skos:Concept.
onto:Equipe rdf:type skos:Concept.
```

Figure 2 : Exemple de représentation du concept *Chercheur* en SKOS

SKOS est formellement défini comme une ontologie OWL (ses éléments sont des classes et des propriétés OWL). Il est simple et intuitif et peut être utilisé seul pour représenter des systèmes simples (les thésaurus par exemple) ou combiné à d'autres langages formels tels qu'OWL pour décrire des modèles plus complexes (une ontologie formelle, par exemple) (Jupp et al., 2008).

Une fois les connaissances décrites dans des langages formels, elles doivent être accessibles et exploitables par des systèmes informatiques. Dans ce sens, des langages dédiés ont été développés pour l'interrogation de ces données structurées. Dans la section suivante, nous décrivons le langage de requêtes le plus couramment utilisé pour manipuler le contenu des bases de données RDF : SPARQL.

5.6 SPARQL (SPARQL Protocol And RDF Query Language)

Comme le langage SQL, SPARQL[10] est un langage de requêtes destiné à interroger les bases de triplets RDF, appelées aussi triplestores. Il est devenu une recommandation du W3C depuis janvier 2008. Il utilise des patterns de graphe pour déterminer les triplets qui satisfont les conditions des requêtes. En principe, avec ce langage, on peut accéder à toute donnée du Web représentée au format RDF. SPARQL utilise une syntaxe inspirée de SQL et est à ce titre très similaire à ce langage. Le schéma général d'une requête SPARQL est de la forme suivante :

[10] http://www.w3.org/TR/rdf-sparql-query/

```
# déclaration de préfixes
PREFIX foo: <http://example.com/resources/> ...
# définition des jeux de données
FROM ...
# clause résultat
SELECT ...
# motif de la requête
WHERE { ... }
# modificateur de requête
GROUP BY
```

Une requête SPARQL permet d'extraire des triplets d'un graphe RDF vérifiant certaines conditions définies dans sa clause ***where***. SPARQL possède également d'autres clauses telles que les opérateurs booléens (union, intersection), de filtrage sur les valeurs pour exprimer des requêtes plus complexes mais plus spécifiques. SPARQL est souvent utilisé conjointement avec des langages de programmation comme Java pour manipuler des données RDF dans des applications.

Mentionnons pour terminer notre description des langages, le projet OBO[11] (Open Biomedical Ontologies) et le langage du même nom associé, initié par un groupe d'ontologistes. Ce projet a établi un ensemble de principes pour guider le développement des ontologies biomédicales (Smith et al., 2007). Son objectif est la création d'ontologies de référence interopérables dans le domaine biomédical. Pour cela, un langage de représentation d'ontologie spécifique a été proposé, similaire au langage OWL, pour supporter les besoins de la communauté. Comme les autres langages présentés ci-dessus, le format OBO supporte la définition des classes, relations et instances, mais également des liens entre les relations.

Grâce à leur capacité d'échange et de partage d'informations au sein de communautés, les ontologies ont connu un grand engouement. Dans le domaine biomédical, le besoin de disposer de vocabulaires communs a motivé différentes communautés à développer des RTO en fonction de leurs besoins. Des initiatives pour unifier ces différentes ressources dans un seul système ont également été proposées (Bodenreider, 2004). Ces ressources présentent un intérêt particulier pour notre travail qui s'intéresse, d'une part, à la construction d'une ontologie spécifique à la maladie d'Alzheimer et, d'autre part, à l'utilisation d'une telle ressource pour guider un système de RI. Nous présentons ainsi des RTO très répandues dans le domaine biomédical dans la section qui suit avant de nous focaliser plus spécifiquement sur celles relatives au domaine du vieillissement cérébral et de la neurologie, plus en lien avec notre domaine d'application.

[11] http://www.obofoundry.org/

6 Les ressources termino-ontologiques du domaine médical

Avec le volume et la diversité des données biomédicales, de nombreuses RTO ont été développées pour répondre aux différents besoins des utilisateurs dont principalement la gestion et le partage de l'information médicale. Dans cette section, nous présentons le thésaurus MeSH, la nomenclature SNOMED (Spackman et Reynoso, 2004) et l'Unified Medical Language System (UMLS) (Bodenreider, 2004).

6.1 Le thésaurus MeSH

Le thésaurus MeSH[12] (Medical Subject Headings) a été développé par la NLM (National Library of Medicine) pour supporter l'indexation des articles et revues biomédicaux de la base bibliographique MEDLINE[13]. Il est constitué d'un ensemble de descripteurs organisés suivant une structure hiérarchique de 16 catégories. Depuis sa création en 1954, il est révisé et mis à jour régulièrement et contient 27 149 descripteurs avec 218 000 entrées (termes) dans sa version actuelle (2014). Les composants principaux du thésaurus sont organisés en trois niveaux : les descripteurs, les concepts et les termes.

Dementia [Descriptor]	
Dementia	[Concept, Preferred]
Dementia	[Term, Preferred]
Dementias	[Term]
Amentia	[Term]
Senile Paranoid Dementia	[Concept, Narrower]
Senile Paranoid Dementia	[Term, Preferred]
Dementias, Senile Paranoid	[Term]
Familial Dementia	[Concept, Narrower]
Familial Dementia	[Term, Preferred]
Dementia, Familial	[Term]

Figure 3 : Exemple du descripteur MeSH Dementia

Les termes (ou « entrées » du vocabulaire) sont les unités de base (du niveau de la langue), et plus précisément des chaînes de caractères constituant le vocabulaire du thésaurus. Un concept regroupe un ou plusieurs termes synonymes dont l'un d'eux représente son terme préféré. Il a aussi un identificateur unique (*ConceptUI*). Un descripteur (*Main Heading*) comprend un ou plusieurs concepts dont un est désigné comme son concept préféré. Il possède également un identificateur unique (*DescriptorUI*). Pour un descripteur, les termes associés à ses différents concepts ne sont pas strictement synonymes. En effet, les concepts associés à un descripteur ne sont pas forcément équivalents même s'ils sont sémantiquement proches ; un concept peut être plus spécifique que le concept préféré ; les termes associés héritent ces mêmes relations et restent ainsi non-équivalents. Mais, dans le cadre de l'indexation, les termes contenus dans un descripteur sont généralement considérés comme équivalents. Par exemple, dans la figure 3, le descripteur *Dementia* comprend un concept

[12] http://www.nlm.nih.gov/mesh/meshhome.html
[13] http://www.ncbi.nlm.nih.gov/pubmed

préféré *Dementia* et deux concepts plus spécifiques (Narrower than), *Senile Paranoid Dementia* et *Familial Dementia*.

MeSH est actuellement disponible dans plusieurs langues (anglais, français, etc.). Sa version française est développée et maintenue par l'Institut National français de la Santé et de la Recherche Médicale (INSERM)[14].

6.2 La nomenclature SNOMED

La SNOMED (Systematized Nomenclature of Medicine) (Spackman et Reynoso, 2004) est la terminologie clinique la plus complète. Elle est pluri-axiale (*Body Structure*, *Organism*, *Substance*, etc.) et couvre tous les champs de la médecine et de la dentisterie humaines, ainsi que la médecine animale.

Une version française de la SNOMED 3.5[15] est disponible depuis 1998 et est mise à jour régulièrement et indépendamment de la SNOMED International (version originale), qui a évolué et été combinée aux Read Codes pour créer la SNOMED CT[16] (SNOMED Clinical Terms). La SNOMED CT est une terminologie clinique représentant l'ensemble des termes médicaux utilisés par les praticiens de santé et permet ainsi l'exploitation des données cliniques. Elle est utilisée aux Etats-Unis pour coder l'ensemble des données du dossier patient. Dans sa version actuelle, elle contient plus de 311 000 concepts, définis formellement et organisés en hiérarchies. Chaque concept possède un identificateur unique (*concept identifier*) et peut avoir plusieurs termes synonymes. Elle est constituée de près d'un million de termes anglais dans sa version internationale. En plus des relations hiérarchiques (définies par *is-a*), les concepts de la SNOMED CT sont associés par d'autres types de relations (*causative_agent*, *finding_site*, etc.).

La SNOMED CT constitue ainsi, pour les systèmes d'information sanitaires, un moyen efficace et fiable pour la représentation et l'accès aux informations cliniques (Pereira et al., 2009; Giannangelo et Fenton, 2008).

6.3 L'UMLS

La ressource la plus importante dans le domaine biomédical reste aujourd'hui l'UMLS (Lindberg et al., 1993; Bodenreider, 2004) qui intègre actuellement plus de 160 RTO, incluant le MeSH, la Gene Ontology (GO) (Ashburner et al., 2000) et la SNOMED CT. Il est composé d'un réseau sémantique, d'un Metathesaurus et d'un lexique spécialisé. Le réseau sémantique de l'UMLS comprend 133 types sémantiques (exemple : *Sign or Symptom*) organisés hiérarchiquement et suivant plus de 50 types de relations associatives. Le Metathesaurus est un large graphe constitué de plus de plus de deux millions de nœuds (concepts) et de plus de 47 millions de relations entre ces concepts. Les concepts sont constitués de termes synonymes provenant des divers vocabulaires sources, présentant ainsi un sens unifié pour l'ensemble des termes. Chaque concept a un identifiant unique, appelé CUI (Concept Unique Identifier), parfois une définition, des synonymes (en anglais et éventuellement dans d'autres langues) et est catégorisé par un ou plusieurs types sémantiques (par exemple, le concept *Alzheimer's*

[14] http://mesh.inserm.fr/mesh/
[15] http://esante.gouv.fr/snomed/snomed/
[16] http://www.ihtsdo.org/snomed-ct/

disease a pour type sémantique *Disease or Syndrome*). Le lexique spécialisé spécifie des informations morphosyntaxiques et orthographiques des termes, qui sont associés (indirectement) aux concepts. En plus de ces trois composants, l'UMLS est associé à un ensemble d'outils facilitant son exploitation. Une présentation plus détaillée de l'UMLS est faite dans le chapitre 3 car nous nous sommes appuyés sur cette ressource pour le travail consacré à la construction d'OntoAD.

Dans la section suivante, nous rappelons rapidement les principales ressources existantes dans le domaine de la neurologie et du vieillissement, domaine en lien direct avec le cas d'application privilégié de notre travail.

6.4 Les ontologies dans le domaine de la neurologie

Concernant les maladies neuro-dégénératives, comme celle d'Alzheimer, des ressources telles que l'ontologie SWAN (Semantic Web Application in Neuromedicine) (Ciccarese et al., 2008) et l'ontologie ND (Neurological Disease Ontology) (Jensen et al., 2013) ont été récemment développées pour supporter les applications du domaine. L'ontologie SWAN (Ciccarese et al., 2008) a été construite grâce à une collaboration entre des chercheurs d'Alzforum[17] (*Alzheimer Research Forum*) et des informaticiens du *Massachusetts General Hospital* et de l'université de Harvard. Elle s'inscrit dans le projet SWAN dont l'objectif est de fournir un cadre pour le stockage, l'accès, l'intégration et le partage des informations scientifiques de façon structurée ou semi-structurée en biomédecine en général et en neurologie en particulier. Elle a été initialement conçue pour modéliser les connaissances scientifiques du domaine et supporter ainsi les applications des chercheurs biomédicaux. Pour faciliter sa réutilisabilité et son intégration avec d'autres ontologies, l'ontologie SWAN a été, par la suite, développée de façon modulaire ; elle est ainsi constituée de plusieurs modules. Le module principal de l'ontologie représente formellement (en OWL) les éléments du discours scientifique. Ce discours est décrit par des concepts (*research statement*, *claim*, *hypothesis*, *comment* et *research question*) et des relations entre ces concepts (*consistent with*, *inconsistent with*, *discuss*, *contain*, *alternative to*). Ce module (nommé *scientific discourse*) représente l'infrastructure de base et peut être intégré à des ontologies de domaine pour des applications spécifiques. Par exemple, le module *swan alzheimer* intègre tous les modules nécessaires pour supporter la base de connaissances de la maladie d'Alzheimer, développée dans le cadre du projet SWAN. La *Gene Ontology* est un exemple d'ontologie intégrée pour l'annotation des gènes.

L'ontologie ND (Jensen et al., 2013) a été conçue récemment pour représenter formellement les connaissances sur les maladies neurologiques en général. Son objectif est de modéliser les aspects cliniques et basiques de ces types de maladies. Ainsi des concepts spécifiques décrivant les maladies neurologiques, comme leurs signes, symptômes, pathologies et traitements associés, sont formalisés. Cette ontologie est également modulaire et étend les ontologies BFO (Basic Formal Ontology) (Grenon et al., 2004) et OGSM (Ontology for General Medical Science) (Scheuermann et al., 2009).

[17] http://www.alzforum.org/

A côté des ressources présentées dans cette section et qui sont principalement en anglais, notons l'existence du portail CISMeF[18] (Catalogue et index des Sites Médicaux français) dont l'objectif est d'indexer les sites médicaux francophones et qui intègre d'importantes RTO de santé en français dont le MeSH, la SNOMED 3.5 et la CIM10.

Si on note de nos jours une prolifération de RTO, leur construction n'est pas une tâche aisée. Cette activité fait l'objet de recherches actives au sein de la communauté de l'IC. Nous faisons dans la section suivante une revue des principales approches proposées pour la construction d'ontologies.

7 Les approches de construction d'ontologies

L'ingénierie d'ontologies consiste en l'étude des méthodes, techniques et outils pour traiter les différentes phases de développement d'une ontologie. Elle s'intéresse notamment aux méthodologies qui fournissent les orientations et techniques pour supporter le processus de construction d'ontologies. Vu l'importance des ontologies, plusieurs approches méthodologiques ont été proposées pour guider ce processus mais il n'en existe pas qui soit acceptée par tous (Lopez, 1999).

On peut distinguer quatre grandes catégories d'approches : i) les approches de construction d'ontologies à partir de zéro ; ii) les approches de construction d'ontologies à partir de textes ; iii) les approches basées sur la réutilisation de RTO existantes ; et enfin iv) les approches basées sur le « crowdsourcing ». La méthodologie varie en fonction de l'usage qu'on veut faire de l'ontologie mais aussi des ressources disponibles à utiliser pour la construction. Les domaines d'applications des ontologies étant variés, les approches restent différentes et peuvent être manuelles ou semi-automatiques. Les premières propositions dans le domaine étaient des méthodes manuelles où les ontologies sont construites à partir de zéro avec l'aide d'experts de domaines. Ces quinze dernières années, l'ingénierie ontologique a été marquée par les approches de construction d'ontologie à partir de textes (Cimiano et Völker, 2005; Buitelaar et al., 2004; Aussenac-Gilles et al., 2008). Ces approches se servent des outils et méthodes du traitement automatique des langues (TAL). Elles visent surtout à alléger le processus de construction d'ontologies en automatisant certaines étapes grâce aux textes qui constituent des sources de connaissances très riches. Le développement des supports électroniques a aussi rendu ces données textuelles plus accessibles et ont ainsi facilité leur exploitation par des outils de TAL. Enfin, face au challenge et compte tenu des efforts requis pour la construction d'ontologies, la réutilisation de RTO existantes est une question très importante. Ainsi, de nouvelles approches se sont intéressées à la réutilisation et la réingénierie de ressources existantes dans l'ingénierie ontologique (Simperl, 2009; Hepp et de Bruijn, 2007; Villazón-Terrazas et Gómez-Pérez, 2012).

Dans cette section, nous exposons les travaux sur la construction d'ontologies selon les quatre approches précédentes. Nous proposons ainsi un survol des méthodes manuelles de construction d'ontologies avant de se focaliser plus particulièrement sur les méthodologies de construction d'ontologies à partir de textes. Nous décrivons par la suite les approches

[18] http://www.hetop.eu/hetop/

s'intéressant à la réutilisation de RTO existantes. Enfin, nous terminons cette partie par un aperçu sur l'utilisation du crowdsoursing en ingénierie ontologique.

7.1 Les méthodologies de construction d'ontologies à partir de zéro

Les premières méthodologies de construction d'ontologies ont été proposées au milieu des années 1990 avec les travaux de (Uschold et King, 1995) et de (Gruninger et Fox, 1996). Ensuite, des méthodologies plus détaillées comme Methontology (Lopez et al., 1997) et celle plus récente de Sure et ses collègues (Sure et al., 2003) sont apparues. Dans ces méthodologies, les sources de connaissances sont généralement construites par des humains et notamment des experts du domaine d'application. Des techniques d'acquisition de connaissances spécifiques (réunions de brainstorming, interview d'experts, etc.) sont nécessaires mais notons que rien ne garantit l'exhaustivité des connaissances décrites.

Nous présentons dans cette sous-section quelques approches considérées comme parmi les plus représentatives.

L'approche proposée dans (Uschold et King, 1995) est basée sur l'expérience acquise lors du développement de l'ontologie *Enterprise Ontology* modélisant les activités de l'entreprise. Elle fournit les lignes directrices pour guider le processus de construction d'ontologie et comprend quatre étapes :

- la phase d'identification du but, de la portée et des utilisations prévues pour l'ontologie ;
- la phase de développement qui consiste à identifier les concepts, les relations du domaine d'intérêt et à les représenter explicitement dans un modèle formel. Cette phase prévoit également l'exploitation éventuelle d'ontologies existantes ;
- la phase d'évaluation visant à analyser l'ontologie résultante pour vérifier si elle satisfait les utilisations prévues ;
- la phase de documentation ayant pour but de faciliter la réutilisation et le partage de l'ontologie résultante.

Dans cette méthodologie, les activités et les techniques utilisées lors des différentes étapes sont peu détaillées. Par exemple, il n'y a pas d'indication sur la manière de déterminer les concepts clefs de l'ontologie.

Une autre méthodologie a été proposée dans (Gruninger et Fox, 1996) et utilisée dans le cadre du projet TOVE (*Toronto Virtual Enterprise*) toujours dans le domaine de l'entreprise. Selon les auteurs, le développement d'une ontologie doit être motivé par des problèmes (scenarii) qui se posent dans le domaine d'application (ici, l'entreprise). Ces derniers sont formulés sous forme de questions informelles auxquelles l'ontologie doit permettre de répondre. Les termes extraits de ces questions sont utilisés pour spécifier la terminologie dans un langage formel. Cette méthodologie a permis de développer des projets complexes dans le domaine de l'entreprise mais reste limitée car ni les différentes étapes ni les techniques utilisées ne sont décrites précisément.

Methontology (Lopez et al., 1997) est une méthodologie de construction d'ontologie développée au Laboratoire d'intelligence artificielle de l'université polytechnique de Madrid. Dans cette approche, le processus de développement d'une ontologie comprend des activités

de gestion de projet (planification, contrôle, assurance de la qualité), des activités orientées développement (spécification, conceptualisation, formalisation, etc.) et des activités de support (évaluation, documentation). Les différentes étapes sont bien identifiées et les activités réalisées dans chaque étape sont décrites. Elle adopte des techniques du génie logiciel dans son processus de développement et couvre ainsi tout le cycle de vie de l'ontologie. Methontology est supportée par la plateforme ODE (Ontology Developpement Environment) qui a donné suite à WebODE. Beaucoup d'ontologies ont été développées en utilisant cette méthodologie : *CHEMICALS* (dans le domaine des éléments chimiques), l'ontologie $(KA)^2$ (connaissances dans la communauté scientifique), la *Reference Ontology*, etc. De plus, c'est une recommandation de la *Foundation for Intelligent Physical Agents* (FIPA).

Dans (Noy et Mcguinness, 2001), les auteurs ont proposé une méthodologie itérative de construction d'ontologies qui comprend sept étapes. La première étape consiste à identifier les besoins et à délimiter le domaine de connaissances à modéliser. La deuxième étape recommande la réutilisation d'éventuelles ontologies existantes et pertinentes pour le domaine. La troisième étape permet d'identifier les termes importants pour l'ontologie cible. Les quatrième et cinquième étapes consistent à définir respectivement les classes de l'ontologie et leur hiérarchie, et les attributs des classes (concepts) et les relations qu'elles entretiennent. Pour construire la hiérarchie entre les concepts de l'ontologie, différentes stratégies peuvent être utilisées (Uschold et Gruninger, 1996) : 1) la méthode *top-down* où l'on commence par la définition des concepts les plus généraux pour ensuite définir leurs sous-concepts; 2) la méthode *bottom-up* qui commence par la définition des concepts les plus spécifiques qui sont ensuite subsumés par des concepts plus généraux ; 3) la combinaison des deux méthodes en définissant d'abord les concepts les plus intéressants. L'étape 6 permet de définir les propriétés associées aux relations telles que leurs domaines et co-domaines tandis que la septième et dernière étape est consacrée à la création des instances. Bien que moins complète que la Methontology (il n'y a pas d'étape de formalisation et d'évaluation), cette méthode est intéressante car elle définit des principes très précis sur les choix de représentation des classes, instances, relations ainsi que leur structuration qui permettent de constituer le noyau d'une ontologie.

Une description plus détaillée de ces méthodologies et d'autres moins courantes ainsi que leur comparaison est proposée dans (Lopez 1999; Mizoguchi, 2004). Bien qu'elles aient été utilisées avec succès pour construire des ontologies, la plupart de ces méthodologies sont manuelles. De plus, les différentes étapes du processus sont souvent trop peu détaillées ; la manière dont les concepts et les relations sont choisis et définis n'est pas précisée.

La construction manuelle d'ontologie étant coûteuse (en temps et ressources) et fastidieuse, d'autres approches proposent d'alléger le processus en automatisant certaines étapes. C'est le cas des méthodes d'acquisition d'ontologie à partir de textes (Buitelaar et al., 2004; Cimiano et Völker, 2005; Biébow et Szulman, 1999; Aussenac-Gilles et al., 2008) qui ont été largement utilisées ces quinze dernières années dans l'ingénierie ontologique.

Dans la section suivante, nous présentons ces approches qui considèrent principalement les corpus de texte comme sources de connaissances.

7.2 Les méthodologies d'acquisition d'ontologies à partir de textes

Avec les avancées en TAL, les méthodes et les outils d'extraction de connaissances à partir de texte sont devenus robustes. L'analyse et le traitement des documents textuels deviennent ainsi plus simples. Ce qui motive le développement de méthodologies de construction d'ontologies à partir des textes, c'est que ces derniers sont considérés comme de bonnes sources de connaissances. Ces approches combinent souvent des techniques linguistiques, statistiques et/ou d'apprentissage automatique pour extraire des connaissances ontologiques à partir de textes (Zouaq et Nkambou, 2010). Les techniques linguistiques procèdent à une analyse superficielle ou profonde des textes tandis que les méthodes statistiques et d'apprentissage automatique exploitent des informations telles que les fréquences documentaires des termes, leurs nombres d'occurrences dans le corpus et leurs cooccurrences. Dans la pratique, ces deux techniques sont souvent combinées. Elles sont généralement appliquées pour la construction d'ontologies de domaine. Une étape préalable est donc la constitution d'un corpus représentatif et adéquat du domaine. Ensuite, des techniques de TAL sont utilisées pour traiter le corpus afin d'extraire certains, voire tous les constituants de l'ontologie : les concepts, les relations, les instances et les axiomes.

Nous présentons dans ce qui suit quelques méthodologies considérées comme représentatives de cette catégorie dans la littérature.

Dans (Maedche et Staab, 2001), une méthode itérative pour l'acquisition semi-automatique d'ontologie mais aussi pour l'enrichissement d'ontologies existantes est proposée. Elle fournit un ensemble d'algorithmes organisés en modules permettant d'extraire des primitives (concepts, relations, etc.) à partir de textes. C'est une méthode qui combine des techniques de TAL, des techniques d'apprentissage automatique et des méthodes statistiques. Pour l'extraction des termes de l'ontologie, une méthode basée sur des mesures statistiques est appliquée aux N-grammes. Une méthode de clustering est ensuite utilisée pour regrouper ces termes en concepts. Concernant l'extraction des relations hiérarchiques, les relations syntaxiques et les relations d'expansion (voir le chapitre 3 de ce présent manuscrit, section 2.3) sont exploitées tandis que la technique des règles d'association permet d'acquérir les relations non hiérarchiques. Dans cette méthode, la conceptualisation est automatique ; elle permet de générer une ontologie automatiquement ; cette dernière peut ensuite être raffinée et enrichie avec l'aide d'un expert (ajout de nouveaux concepts pertinents, suppression de concepts non pertinents). Cette méthode a été implémentée dans l'outil de construction d'ontologies Text-To-Onto (Maedche et al., 2000).

Cimiano et ses collègues (Cimiano et Völker, 2005) ont développé une autre approche pour supporter l'acquisition automatique d'ontologies à partir de textes. Elle combine des méthodes linguistiques (segmentation, lemmatisation, analyse grammaticale, etc.) et des techniques d'apprentissage automatique. L'analyse linguistique permet d'extraire les éléments constitutifs de l'ontologie tandis que les algorithmes d'apprentissage servent à calculer leur pertinence. C'est une approche itérative où l'ontologie peut être raffinée et enrichie avant sa validation par des experts. Un outil nommé Text2Onto[19] a été développé pour supporter cette méthodologie. Il consiste en une suite de modules développés en Java qui permettent

[19] https://code.google.com/p/text2onto/

d'extraire les primitives à partir de textes ; chaque module peut combiner un ou plusieurs algorithmes : RTF (Relative Term Frequency), TF.IDF (Term Frequency – Inverse Document Frequency) (Salton et al., 1975), l'entropie, ou la méthode C-value/NC-value proposée dans (Frantzi et al., 1998) pour l'extraction des concepts ; exploitation de WordNet[20](Fellbaum, 1998), et définition de patrons lexico-syntaxiques (Hearst, 1992) pour l'extraction des relations de subsomption ; utilisation de WordNet et définition des expressions régulières avec JAPE[21] pour l'extraction des relations partitives ; analyse syntaxique pour déterminer les relations sémantiques générales. Text2Onto est fondée sur l'architecture GATE[22] et est très flexible et donc extensible : de nouveaux algorithmes peuvent être intégrés facilement. Il intègre une fonctionnalité permettant de tenir compte des changements dans le corpus et permet ainsi de suivre l'évolution de l'ontologie en fonction des changements dans le corpus. L'outil dispose d'une interface graphique utilisateur avec plusieurs volets et est indépendant du langage de formalisation utilisé. Text2Onto est intégré dans l'environnement d'ingénierie ontologique NeOn[23] (Suarez-Figueroa et Gomez-Perez, 2009). Il est également disponible comme un plug-in de l'environnement de développement Eclipse[24], ce qui facilite son intégration dans d'autres éditeurs d'ontologies.

Terminae[25], développé au LIPN par le groupe de recherche *Terminologie et Intelligence Artificielle* (TIA), qui est à la fois une méthodologie et un outil de construction semi-automatique d'ontologies à partir de textes (Biébow et Szulman, 1999) a également été proposé comme alternative. Il s'appuie sur des méthodes et des outils de TAL pour extraire les éléments de l'ontologie et ceux de l'ingénierie de connaissances pour la représenter. L'approche a été élaborée sur la base d'expériences pratiques dans plusieurs domaines. Elle couvre l'ensemble du processus de développement de l'ontologie qui comprend quatre phases: la constitution du corpus (documents techniques, comptes rendus, articles scientifiques, etc.), l'étude linguistique (identification des termes et de leurs relations), la normalisation ou conceptualisation (concepts et relations désambigüisés) et la formalisation. La phase de conceptualisation est manuelle et guidée; l'ingénieur ontologique est chargé de regrouper les termes extraits en concepts. L'outil Terminae utilise un formalisme des logiques de description pour représenter l'ontologie. Il peut intégrer les résultats de différents outils d'extraction terminologique, tels que Syntex (Bourigault et Fabre, 2000) et YateA (Aubin et Hamon, 2006). En plus, il permet le traitement de corpus multilingues (français et anglais actuellement).

Buitelaar et ses collègues (Buitelaar et al., 2004) ont proposé une méthode principalement basée sur la linguistique. Elle définit des règles linguistiques qui permettent d'extraire des concepts et des relations à partir de collections de textes annotés linguistiquement. C'est une approche qui intègre l'analyse linguistique dans l'ingénierie ontologique. Elle supporte l'acquisition semi-automatique et interactive d'ontologies à partir de textes mais aussi

[20] http://wordnet.princeton.edu/
[21] http://gate.ac.uk/sale/tao/splitch8.html
[22] https://gate.ac.uk/
[23] neon-toolkit.org/
[24] http://www.eclipse.org
[25] http://lipn.univ-paris13.fr/terminae/index.php/Main_Page

l'extension d'ontologies existantes. Cette méthodologie est associée à un plug-in OntoLT[26] pour Protégé[27] (Knublauch et al., 2004) qui reste l'outil d'édition d'ontologies le plus utilisé de nos jours. OntoLT utilise des règles de correspondance prédéfinies qui permettent d'extraire automatiquement des classes et des relations candidates dans des textes. Par exemple, une règle fait correspondre le nom en tête de l'expansion à une classe *C* de l'ontologie et le nom avec son modificateur à une sous-classe de *C* permettant d'établir ainsi des relations du type *déclin cognitif léger* est sous-classe de *déclin cognitif.* Une autre règle fait correspondre le sujet à une classe, le prédicat à une relation, le complément d'objet à une classe et crée la relation associative correspondante entre les deux classes. Si une règle est satisfaite, les opérateurs correspondants sont activés pour créer des classes, des relations ou mêmes des instances qui seront par la suite validées et intégrées dans l'ontologie. L'ontologie extraite est intégrée et peut être explorée dans l'environnement de développement Protégé (Knublauch et al., 2004), ce qui facilite la gestion et le partage des ontologies résultantes. Cette approche a été utilisée pour construire une ontologie dans le domaine de la neurologie.

Une autre méthodologie a été proposée et implémenté dans le système OntoGen (Fortuna et al., 2006; Fortuna et al., 2007) pour la construction semi-automatique d'ontologies. OntoGen vise à aider l'utilisateur (souvent un expert du domaine) à identifier les primitives de l'ontologie à partir d'une collection de documents. Pour cela, il s'appuie sur les techniques d'analyse sémantique latente (LSA) (Deerwester et al., 1990b) et de clustering (k-means) (Jain et al., 1999) pour suggérer des concepts, des relations entre ces concepts et des instances. Il offre une interface graphique avec plusieurs volets qui permettent à l'utilisateur (qui peut être un non professionnel de l'IC) de visualiser et d'explorer les concepts mais aussi d'ajuster l'ontologie en ajoutant de nouveaux concepts ou en éditant ceux déjà existants.

Une méthodologie implémentée dans l'outil OntoLearn (Velardi et al., 2006) fournit différentes techniques pour extraire des connaissances ontologiques à partir de textes. Pour l'extraction des termes pertinents d'un domaine, des outils linguistiques et statistiques sont combinés afin de déterminer leur distribution dans le corpus. Elle se sert également de glossaires disponibles sur le Web. Des patrons lexico-syntaxiques décrits par des expressions régulières sont utilisés pour découvrir les relations de subsomption entre concepts. La structure interne des termes multi-mots est aussi utilisée pour extraire ce type de relations, comme dans (Buitelaar et al., 2004). L'utilisation de la base de données lexicale WordNet permet également d'extraire des synonymes et d'autres types de relations.

Comme on peut le constater, ces méthodologies de construction d'ontologies à partir de textes sont basées principalement sur des outils TAL et utilisent souvent des méthodes statistiques pour le filtrage des résultats. Une des limites principales qu'on peut noter pour ces méthodes est qu'elles peuvent fonctionner pour des tâches de construction où un corpus textuel suffisant est disponible pour faire fonctionner les outils de TAL.

Pour une description plus détaillée et une comparaison de ces différentes méthodologies, le lecteur pourra se référer à la synthèse faite dans (Buitelaar et Magnini, 2005) et celle plus récente de Zouaq et Nkambou (Zouaq et Nkambou, 2010).

[26] http://olp.dfki.de/OntoLT/OntoLT.htm
[27] http://protege.stanford.edu/

Dans la section suivante, nous décrivons d'autres approches de construction d'ontologies qui, pour alléger le processus, proposent la réutilisation et la réingénierie de RTO existantes.

7.3 Les approches basées sur la réutilisation de ressources termino-ontologiques existantes

Pour simplifier les tâches de construction d'ontologie, des approches alternatives aux précédentes ont préconisé la réutilisation de RTO existantes (Simperl, 2009). Ces approchent s'intéressent à l'exploitation de l'ensemble ou d'une partie des informations contenues dans ces sources souvent informelles pour le développement de nouvelles ontologies ou l'enrichissement d'ontologies existantes. Elles cherchent souvent à capturer les connaissances implicites contenues dans ces ressources, généralement représentées dans des langages informels (Hepp et de Bruijn, 2007), pour les décrire dans un modèle formel.

Concernant la réutilisation d'ontologies formelles, (Farquhar et al., 1997) ont proposé un ensemble d'outils pour supporter la construction collaborative d'ontologies communes partagées par des groupes distants. Leur objectif est de permettre aux utilisateurs de créer, publier et éditer des ontologies sur un serveur, comme *Ontolingua*, et de les utiliser à grande échelle.

Maedche et ses collègues (Maedche et al., 2003) considèrent la réutilisation et l'évolution des ontologies distribuées comme une partie intégrante de l'ingénierie ontologique. Ils proposent pour cela la mise en place de registres pour localiser les ontologies existantes, et des techniques pour supporter la réutilisation et l'évolution des ontologies distribuées. La localisation des ontologies doit être assurée par leurs descriptions dans une méta-ontologie (exemples: auteur, emplacement, langages utilisés, version, etc.).

7.3.1 Principe suivi par ces approches

Une méthodologie générale proposée dans (Villazón-Terrazas et Gómez-Pérez, 2012) définit des techniques de réutilisation de ressources ontologiques et non-ontologiques pour faciliter le développement de nouvelles ontologies. Elle fournit une procédure pour convertir ces ressources de connaissances en ontologies en utilisant la large base de données lexicale WordNet pour rendre explicites les relations extraites des sources.

Dans (Lonsdale et al., 2010), les auteurs proposent une méthodologie générique et automatique basée sur la réutilisation d'ontologies existantes. La procédure prend en entrée des documents textuels spécifiques au domaine cible et les ontologies sources (Figure 4). Dans cette architecture, le processus de construction d'une ontologie comprend trois étapes : la sélection des concepts, l'extraction des relations et la spécification des contraintes. La phase de sélection de concepts consiste à identifier, à partir des ontologies sources, les concepts importants pour le domaine en se basant sur les documents. Ainsi, les documents doivent être représentatifs et contenir suffisamment d'informations pour couvrir les concepts du domaine. Un algorithme simple de repérage de concepts dans les documents permet de sélectionner les concepts d'intérêt. La deuxième phase permet de déterminer les relations entre les concepts extraits dans l'étape précédente. Ces relations peuvent être déduites automatiquement à partir des ontologies sources et éventuellement à partir des documents. La dernière étape permet de spécifier les contraintes sur les concepts et les relations extraits en

utilisant les ontologies sources et éventuellement les documents. A partir de ces trois composants, une ontologie est générée automatiquement et est ensuite validée par des experts du domaine.

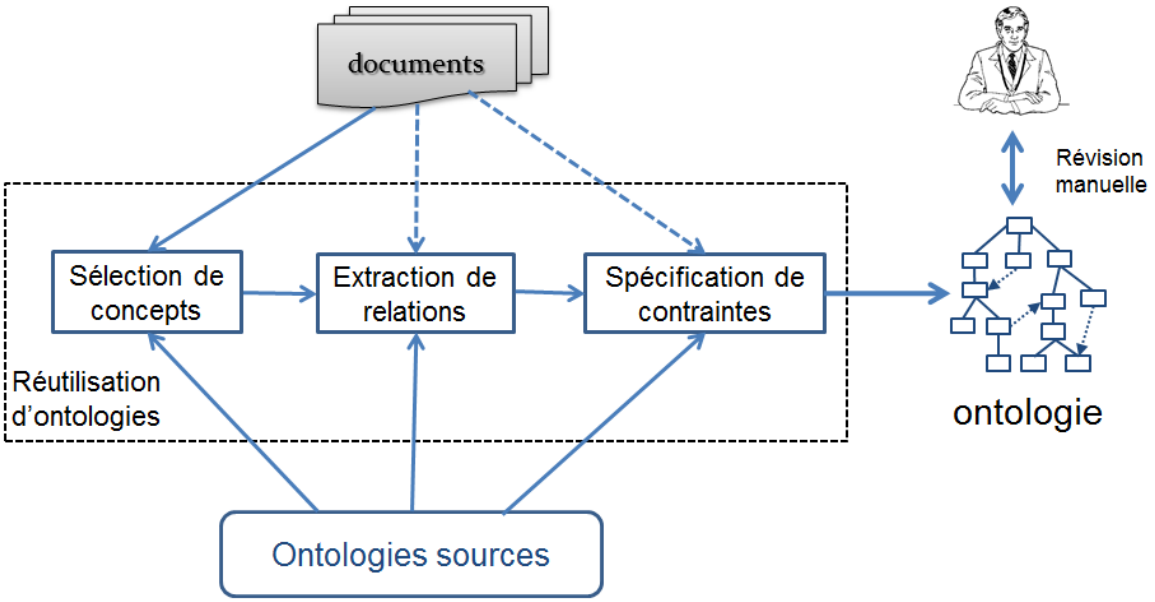


Figure 4 : Architecture générique pour la réutilisation d'ontologies (Lonsdale et al., 2010)

Dans (Hepp et de Bruijn, 2007), les auteurs ont également proposé une méthodologie générique et semi-automatique pour transformer des schémas de classification, des taxonomies informelles et des thésaurus en des ontologies formelles légères. Dans leur approche, l'intervention humaine est limitée à la validation de l'ontologie générée, mais les hiérarchies des terminologies sources sont préservées et considérées comme les relations taxonomiques pour l'ontologie cible bien que ces dernières ne soient pas toujours de vraies relations de subsomption.

Chrisment et ses collègues (Chrisment et al., 2008) ont quant à eux proposé une méthode combinant un thésaurus et un corpus textuel pour construire une ontologie dans le domaine de l'astronomie. Pour la hiérarchie des concepts, ils considèrent les relations *<est plus générale>* et *<est plus spécifique>* qui sont plus vagues que les relations taxonomiques (définies par la relation *is_a*) ; ces relations sont considérées systématiquement comme des relations taxonomiques. En outre, ils ne fournissent pas de mécanismes automatiques pour expliciter ces relations.

7.3.2 Réutilisation de ressources sémantiques dans le domaine biomédical

Dans le domaine biomédical, Hahn et Schulz (Hahn et Schulz, 2004) ont proposé une méthodologie basée sur le Metathesaurus de l'UMLS pour développer une ontologie formelle. La méthode se sert d'une partie (l'anatomie et les pathologies) des connaissances contenues dans cette abondante ressource pour générer un modèle formel. Ils se sont intéressés à la modélisation des relations taxonomiques et partitives pour qu'elles supportent la déduction de

nouvelles connaissances. Pour cela, les relations de généralisation/spécialisation sont explorées pour distinguer manuellement les relations taxonomiques des relations partitives. Cette approche comprend quatre étapes : 1) l'extraction automatique des concepts et relations hiérarchiques (*part_of*, *is_a*) à partir de l'UMLS, 2) la vérification de la consistance, 3) la correction manuelle des cycles et inconsistances et 4) la validation par un expert du domaine. Elle nécessite donc une intervention humaine pour analyser les inconsistances mais aussi une validation manuelle de ces différentes relations.

Assem et ses collègues (Assem et al., 2004) ont proposé à leur tour une méthodologie basée sur la réutilisation de thésaurus pour générer des ontologies. Leur méthodologie comprend quatre étapes : la phase d'analyse de la ressource, la phase de conversion syntaxique, la phase de conversion sémantique et la phase de standardisation. La première étape consiste en l'étude de la ressource et de son contenu. La deuxième traite les aspects syntaxiques liés à la conversion tandis que la troisième étape s'intéresse aux aspects sémantiques. La dernière étape permet la représentation des connaissances dans un langage standard comme OWL après leur transformation. Dans cette méthodologie, les concepts du thésaurus (descripteurs MeSH ou synsets WordNet) sont convertis en classes dans l'ontologie. Les relations de généralisation/spécialisation (les relations *RN* et *RB* dans MeSH) ou d'hyponymie (WordNet) de la source, quant à elles, servent de relations taxonomiques dans l'ontologie. Les autres types de relations sont conservés et représentés en OWL. Cette approche a été expérimentée sur le thésaurus MeSH et sur la base de données lexicale WordNet.

Dans (Bontas et al., 2005), les auteurs ont expérimenté la réutilisation d'ontologies dans les domaines de la médecine et du « eRecruitment » et discuté les challenges soulevés par cette question. Ils montrent, à travers leurs expérimentations, que les avantages d'une approche basée sur la réutilisation dépendent de la nature des sources réutilisées mais aussi du domaine d'application. Elle peut être bénéfique par exemple dans le domaine médical où les RTO sont abondantes. Cette méthode inclut une phase d'enrichissement où de nouvelles connaissances non contenues dans les sources sont intégrées manuellement dans la nouvelle ontologie.

Dans (Jiménez-Ruiz et al., 2008), les auteurs ont également implémenté une méthodologie basée sur la réutilisation d'ontologies. Cette approche extrait simplement les parties pertinentes des ontologies sources pour ensuite les fusionner et les intégrer dans l'ontologie cible. Pour extraire les fragments pertinents à partir de sources de connaissances, un ensemble de concepts de référence est souvent utilisé et aligné aux concepts des sources. Les entités extraites sont ensuite affinées (c'est à dire que des sous-concepts sont définis) ou généralisées manuellement par le développeur de l'ontologie en tenant compte de la cohérence de l'ontologie. Dans ce travail, on suppose que les ontologies à réutiliser sont cohérentes et par conséquent aucun mécanisme pour traiter les incohérences n'est envisagé. Les auteurs l'ont expérimentée en utilisant le thésaurus NCI (National Cancer Institute) et l'ontologie GALEN pour la construction d'une ontologie du domaine de l'arthrite chronique juvénile).

Plus récemment, Charlet et ses collègues ont décrit une autre approche de construction de RTO basée sur la réutilisation et l'intégration de ressources existantes (Charlet et al., 2012). Ils proposent une méthode combinant l'analyse de corpus et la réutilisation de RTO existantes pour construire un noyau ontologique qui est, ensuite, enrichi de manière semi-automatique

avec ces mêmes ressources ; une fois un concept aligné à des concepts des ressources externes, les termes associés à ces derniers sont aussi extraits pour enrichir lexicalement le concept. Cette approche inclut également une phase d'enrichissement semi-automatique avec l'intégration de nouveaux termes et concepts, extraits dans le corpus. Elle a été appliquée en utilisant la Classification Internationale des Maladies[28] (CIM-10) de l'Organisation Mondiale de la Santé, la CCAM (Classification Commune des Actes Médicaux), le FMA, et la SNOMED 3.5 comme terminologies pour développer l'ontologie *ONTOLURGENCES*.

7.3.3 Réutilisation d'ontologies à l'échelle du Web

Concernant la réutilisation à grande échelle, (D'Aquin et al., 2008) considèrent le Web sémantique comme une source de connaissances précieuse pour la construction d'ontologies. Les auteurs proposent ainsi une méthodologie pour la localisation et la réutilisation d'ontologies pertinentes à partir du Web. Un plugin du moteur de recherche de ressources sémantiques, Watson[29] (D'Aquin et Motta, 2011) a été intégré à l'environnement de développement d'ontologies NeOn pour supporter cette approche. En effet, Watson fournit les fonctionnalités nécessaires pour retrouver et explorer les ressources sémantiques publiées en ligne. Cette méthodologie supporte ainsi l'exploitation et la réutilisation à large échelle des ontologies disponibles sur le Web dans l'ingénierie ontologique. En plus, elle permet d'intégrer et de fusionner de manière interactive des connaissances extraites à partir d'ontologies hétérogènes et distribuées.

Alani, quant à lui, s'est intéressé également à la réutilisation d'ontologies à l'échelle du Web (Alani, 2006). Il propose une approche générique de construction d'ontologies combinant des techniques de recherche, de partitionnement, d'alignement et de fusion d'ontologies disponibles en ligne. A partir d'une liste de termes, son système permet de retrouver et d'ordonner les ontologies potentiellement pertinentes en utilisant un outil de recherche d'ontologies comme Swoogle (Ding et al., 2004). Ces dernières sont ensuite analysées et partitionnées pour extraire les parties pertinentes. Ces parties sont, par la suite, comparées et fusionnées pour générer la nouvelle ontologie. Enfin, l'ontologie résultante peut être éditée et raffinée par l'ontologue. Ce travail a permis de mettre en lumière de nombreux challenges soulevés par cette méthode tels que la recherche, le partitionnement, l'alignement et la fusion d'ontologies.

7.3.4 Synthèse sur ces approches de réutilisation

Puisqu'il existe de nombreuses ressources déjà développées, ces approches s'intéressent à l'exploitation de ces ressources pour faciliter et accélérer le processus de construction d'ontologies. Elles utilisent généralement les concepts des sources réutilisées pour générer les classes de l'ontologie (Hahn et Schulz, 2004 ; Chrisment et al., 2008). Concernant la hiérarchie des concepts, certaines méthodes conservent celles issues des sources de connaissances qui sont souvent plus générales que la relation de subsomption (Hepp et de Bruijn, 2007; Chrisment et al., 2008), tandis d'autres se servent de ressources externes pour expliciter ces relations (Villazón-Terrazas et Gómez-Pérez, 2012) ou les raffiner manuellement (Jiménez-Ruiz et al., 2008). Les relations associatives entre ces concepts ne

[28] www.who.int/classifications/icd/en/
[29] http://watson.kmi.open.ac.uk

sont généralement pas prises en compte (Hepp et de Bruijn, 2007) ou sont traitées manuellement (Hahn et Schulz, 2004). Dans (Villazón-Terrazas et Gómez-Pérez, 2012), les auteurs exploitent également des ressources externes telles que WordNet pour désambiguïser les relations associatives. Les ontologies résultantes sont généralement des ontologies légères utilisées principalement pour la RI sémantique (Assem et al., 2004; Soergel et al., 2004; Chrisment et al., 2008; Charlet et al., 2012).

En plus des différentes méthodologies et techniques présentées ici, le développement de larges catalogues pour le stockage, l'accès et l'alignement d'ontologies facilite leur partage et leur réutilisation (Ding et Fensel, 2001; D' Aquin et Noy, 2012). En effet, les catalogues collectent des ontologies à partir de différentes sources de connaissances et proposent des mécanismes pour les retrouver, les explorer et les aligner (Smith et al., 2007; Noy et al., 2009; D'Aquin et Lewen, 2009; Diallo, 2011). Elles facilitent ainsi la publication et la réutilisation d'ontologies existantes pour en générer de nouvelles. Pour une description plus détaillée et une comparaison des bibliothèques d'ontologies, le lecteur peut se référer à la revue faite dans (D'Aquin et Noy, 2012).

7.4 Les approches basées sur le « crowdsourcing »

Contrairement aux approches présentées dans les deux précédentes sections, de nouvelles approches utilisant de plus en plus les techniques de *crowdsourcing* (Sarasua et al., 2012; Mortensen et al., 2013; Getman et Karasiuk, 2014) sont apparues récemment. Le « crowdsourcing » consiste à externaliser des tâches traditionnellement effectuées par un agent désigné (comme un employé ou un entrepreneur) en faisant appel à l'intelligence et au savoir-faire d'un grand nombre de personnes (Howe, 2008). Puisque le processus de construction d'ontologies est fastidieux et nécessite en général beaucoup de temps et de ressources, des chercheurs ont eu l'idée d'impliquer un large groupe d'utilisateurs pour en simplifier la tâche.

(Mortensen et al., 2013) considèrent que le « crowdsourcing » peut être un moyen d'alléger les difficultés soulevées par le développement d'ontologies larges et complexes. Les auteurs ont ainsi proposé des méthodes basées sur le « crowdsourcing » pour réaliser automatiquement diverses tâches de l'ingénierie ontologique, telles que l'évaluation de la qualité de l'ontologie et la production de nouvelles ontologies. Dans cette approche, les participants sont soumis à un test de qualification et ceux qui passent ce test peuvent accéder aux tâches. Les réponses sont ensuite collectées et évaluées. Ils montrent dans leur évaluation que leur méthode pour la tâche de vérification de la hiérarchie d'ontologies donne une précision de 82%.

(Getman et Karasiuk, 2014) ont proposé de manière similaire une méthode reposant sur le « crowdsourcing » pour la construction d'une ontologie dans le domaine du droit. Cette méthode a été implémentée et mise à la disposition d'un groupe de 20 utilisateurs (étudiants en droit) pendant un semestre. Pour simplifier la tâche, chaque utilisateur a travaillé sur un sous-domaine. Après l'évaluation de deux branches (334 concepts structurés par 338 relations sémantiques) de l'ontologie résultante (6000 concepts) par des experts du domaine, ils estiment la couverture des concepts pour les branches analysées à plus de 90%. Bien que leurs résultats fussent concluants, certains problèmes ont été soulevés après analyse. Ils ont noté par exemple que des branches de l'ontologie créées par des utilisateurs différents sont peu connectées. Ils ont remarqué également que des concepts synonymes ont été créés

séparément. Les auteurs concluent que l'implication d'utilisateurs qualifiés dans le « crowdsourcing » est bénéfique pour la construction et la maintenance d'ontologie dans un domaine spécifique comme le droit mais qu'il faut y ajouter une tâche de « polissage ».

En plus de la construction d'ontologies, le « crowdsourcing » a également été utilisé pour l'enrichissement (Lin et Davis, 2010) et l'alignement d'ontologies (Sarasua et al., 2012). Par exemple, dans (Lin et al., 2010), une approche d'évolution d'ontologies basée sur le « crowdsourcing » est décrite. Un système nommé OntoAssist a été développé et utilisé pour intégrer de nouveaux concepts et des relations sémantiques entre eux dans une ontologie.

Ces différents travaux ont montré le potentiel des techniques du « crowdsourcing » dans l'ingénierie ontologique. L'utilisation des connaissances d'un grand nombre d'utilisateurs qualifiés, grâce ces techniques, peut donc être complémentaire aux méthodologies classiques de construction en permettant d'alléger certaines tâches. Cette distribution de tâches fait apparaître la notion de mutualisation et donc de collaboration qu'on retrouve dans une approche émergente, la construction collaborative d'ontologies à travers le Web, notamment matérialisée à travers la version collaborative de l'éditeur Protégé, utilisée pour la construction de la nouvelle version de la CIM : la CIM-11 (Tudorache et al., 2013).

8 Conclusion

Dans ce chapitre, nous avons présenté les différentes notions liées aux ontologies avant de nous focaliser plus particulièrement sur les différentes approches méthodologiques de construction d'ontologies. Les premières méthodes de construction d'ontologies étaient manuelles. Leur application pour la mise en œuvre d'ontologies est fastidieuse et coûteuse. Elles nécessitent beaucoup de temps et de ressources (intervention d'experts de domaine et d'ingénieur de connaissances) pour la construction d'ontologies de taille conséquente. C'est ce qui a motivé le développement des approches d'élaboration d'ontologies semi-automatiques. Parmi ces dernières, les méthodologies de construction d'ontologies à partir de textes ont constitué une des principales propositions. Ces dernières considèrent les textes comme des sources de connaissances précieuses et s'appuient sur des méthodes et outils de TAL pour déterminer les composants de l'ontologie (concepts, relations, axiomes). Pour extraire ces différents éléments à partir des textes, des techniques linguistiques et/ou statistiques sont utilisées indépendamment ou sont combinées. Bien que cette approche de construction d'ontologies ait été appliquée avec succès dans beaucoup de domaines, elle reste confrontée à certaines limites. En particulier, la phase de conceptualisation est soit manuelle, soit automatisée. Dans le premier cas, les mêmes problèmes que les méthodes de construction manuelles se posent tandis que si la conceptualisation est automatique, l'ontologie résultante nécessite des post-traitements afin de ne pas être bruitée. Par exemple, Cimiano et Völker (Cimiano et Völker, 2005), après une comparaison d'une ontologie générée avec Text2Onto et une taxonomie de référence sur le tourisme, ont rapporté une f-mesure de 21,8% (une précision de 17,4% et un rappel de 29,9%) pour la tâche d'extraction de liens taxonomiques. De plus, cette approche ne tire pas profit des ressources sémantiques déjà développées (et abondantes dans certains domaines, comme la médecine) pour faciliter et accélérer le processus de construction d'ontologies. Ainsi, face au challenge de construction d'ontologies

et les efforts requis pour cette tâche, d'autres approches proposent la réutilisation et la réingénierie des RTO existantes. L'idée est d'éviter de « réinventer la roue » en exploitant partiellement ou entièrement les contenus de ces sources de connaissances souvent implicites et informelles pour générer une ontologie. Ces approches ont été largement explorées mais il reste des challenges liés à l'explicitation, la formalisation et l'incomplétude de ces contenus. D'autres travaux ont aussi étudié la combinaison de ces deux approches semi-automatiques.

Dans le chapitre 3, nous proposons une démarche semi-automatique de construction d'ontologies dans le domaine biomédical, qui s'inscrit dans le courant de construction à partir de textes, et fournit différentes techniques pour traiter les problèmes soulevés. Cette méthodologie a été appliquée au domaine médical pour développer l'ontologie OntoAD qui est désormais utilisée pour supporter un modèle de RI.
Dans le prochain chapitre, nous présentons les méthodes de RI qui s'intéressent de plus en plus à l'exploitation des ressources sémantiques, telles que les ontologies, pour améliorer leurs performances.

Chapitre 2: Etat de l'art sur la recherche d'information sémantique

1 Introduction

La RI est un ensemble de méthodes et techniques visant à faciliter l'accès à l'information (Baeza-Yates et Ribeiro-Neto, 1999). L'objectif d'un SRI est ainsi de fournir à ses utilisateurs, à partir d'une collection de documents, ceux qui satisfont leur besoin en information exprimé au travers d'une requête. Pour ce faire, le SRI s'appuie sur un modèle qui définit l'ensemble des fonctions nécessaires pour la description des documents et des requêtes mais aussi les techniques pour leur appariement. En pratique, cela implique donc de multiples tâches : l'acquisition, la représentation, le stockage et l'accès à l'information.

Dans ce chapitre, après une brève présentation des notions et concepts liés à la RI, nous nous intéressons plus particulièrement à la RI dite « sémantique ». D'abord, nous décrivons le principe de fonctionnement d'un SRI. Ensuite, les modèles classiques de RI et les méthodes d'évaluation d'un SRI sont décrits respectivement dans les sections 3 et 4. Dans la section 5, nous nous focalisons sur les différentes approches mettant en œuvre une RI sémantique avant de présenter des travaux de RI sémantique ayant proposé une application au domaine médical (section 6).

2 Fonctionnement d'un système de recherche d'information

Comme nous l'avons dit, le principe d'un SRI est de fournir aux utilisateurs les documents dits pertinents correspondant à leurs besoins. La pertinence est mesurée à partir du degré de correspondance entre la requête et chaque document. L'utilisateur exprime son besoin en information sous forme de requête qu'il soumet au SRI et ce dernier lui retourne les documents jugés pertinents par rapport à la requête formulée. Pour ce faire, le SRI compare la requête aux documents disponibles pour y répondre. Pour permettre cette comparaison, les requêtes et les documents doivent être représentés de manière similaire. Ce processus, comme décrit dans la figure 5, comprend une phase d'indexation, une phase de recherche et une phase d'appariement.

2.1 Indexation des documents

Elle consiste à décrire le contenu des documents de la collection (l'ensemble des documents de l'espace de recherche) par des éléments clés appelés *entrée d'index* ou *descripteurs* ou *termes d'indexation*. Un document correspond à un item de la collection et peut être de différentes natures (texte, image, audio, vidéo, etc.). Il peut s'agir d'un document dans sa totalité ou d'une partie d'un document. Nous nous intéressons dans ce chapitre plus particulièrement aux documents de nature textuelle.

L'indexation consiste à traiter un document afin d'identifier un ensemble de descripteurs significatifs représentant son contenu. Ces descripteurs peuvent être des mots, des n-grammes

(séquence de n mots), des concepts d'un thésaurus ou d'une ontologie (on parle alors d'« indexation conceptuelle » comme nous le décrivons en section 4), etc. L'importance (on parle alors de « poids ») d'un descripteur dans un document dépend généralement du modèle utilisé. Pour l'indexation textuelle, des facteurs tels que la fréquence du terme dans le document (TF), la fréquence du terme dans la collection (DF), la taille du document ou encore la position du terme dans le document sont souvent combinés. La technique consistant à déterminer, pour chaque terme d'indexation, son poids dans un document est appelé la « pondération ». Un schéma de pondération combine généralement plusieurs propriétés de ces descripteurs pour le calcul de leurs poids.

En fonction des ressources utilisées, on peut distinguer deux types d'indexation :

- l'indexation libre où les descripteurs sont choisis librement sans utiliser une liste prédéfinie de termes ;
- l'indexation contrôlée où les descripteurs sont prédéfinis dans une ressource telle qu'une terminologie, un thésaurus ou une ontologie.

En fonction du niveau d'automatisation, on peut également distinguer trois niveaux d'indexation:

- l'indexation manuelle où l'identification des descripteurs considérés comme significatifs pour représenter un document est réalisée manuellement par un spécialiste du domaine d'intérêt ou un documentaliste ;
- l'indexation semi-automatique où l'identification automatique des descripteurs est suivie d'une étape manuelle de vérification et de validation. On parle aussi d'indexation supervisée ;
- l'indexation automatique qui se réalise sans aucune intervention manuelle. Elle est particulièrement adaptée, voire absolument nécessaire, pour le traitement des données volumineuses, comme c'est le cas notamment à l'échelle du Web.

2.2 Recherche de documents

Dans la phase de recherche, l'utilisateur exprime son besoin en information par une requête. Cette dernière peut être formulée de plusieurs manières : par une liste de mots clés, par des expressions booléennes, en langage naturel, etc. Comme les documents, la requête est, elle aussi, transformée et représentée par un ensemble de descripteurs. Elle peut également être reformulée ou étendue pour mieux exprimer le besoin en information de l'utilisateur (Díaz-Galiano et al., 2009). La reformulation consiste à générer une nouvelle requête, censée être plus appropriée, à partir de la requête initiale de l'utilisateur. L'expansion permet, quant à elle, de compléter la requête initiale par des termes d'indexation supplémentaires afin de mieux exprimer le besoin en information de l'utilisateur ; c'est donc une sorte de reformulation de requêtes.

2.3 Appariement documents-requête

Cette étape consiste à estimer la pertinence de chaque document de la collection par rapport à la requête de l'utilisateur afin de classer les documents. Par exemple, en RI textuelle classique, cette correspondance est souvent basée sur les mots que partagent la requête et les documents. Les documents les plus pertinents (on parle des « top» documents) sont ainsi retournés à l'utilisateur. L'appariement des documents à la requête et leur ordonnancement par pertinence constituent une étape cruciale du processus et dépendent du modèle de RI utilisé.

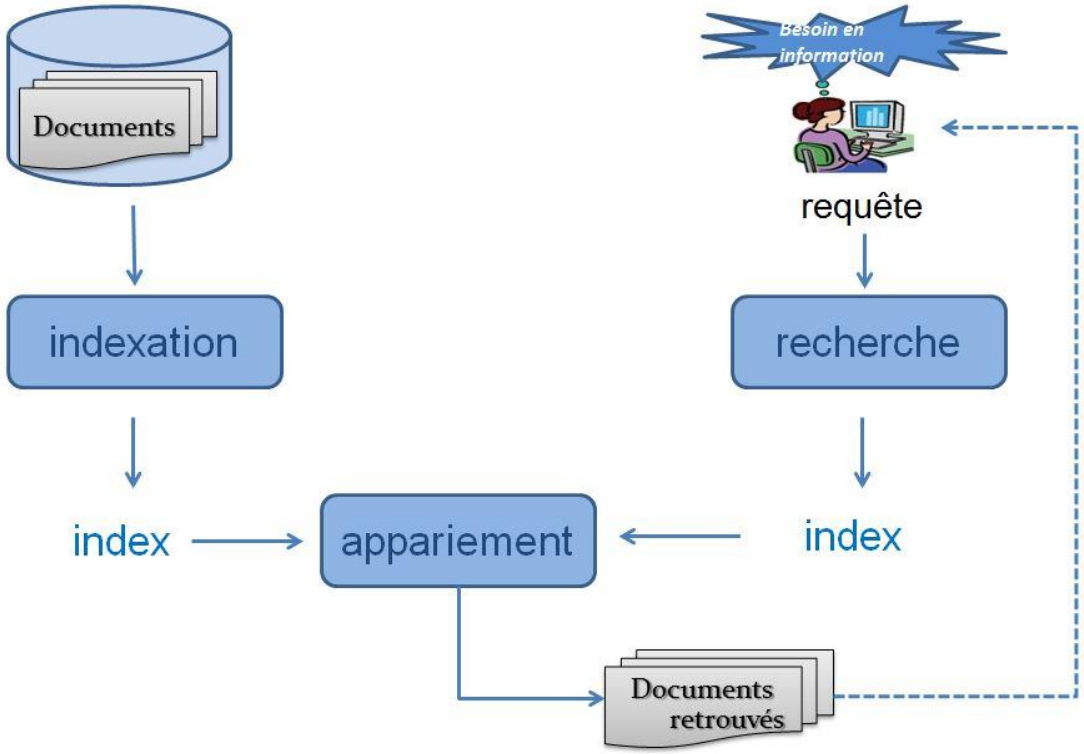


Figure 5 : Processus d'un SRI

Dans la section suivante, nous nous intéressons à ces différents modèles utilisés pour supporter un SRI.

3 Les différents modèles de recherche d'information

Dans la phase d'indexation, chaque document (tout comme les requêtes dans la phase d'interrogation) est représenté par un ensemble de descripteurs avec des poids associés qui déterminent leur importance par rapport à ce document. Un modèle de RI permet l'interprétation de ces poids et fournit le mécanisme pour déterminer le degré de correspondance entre les documents et la requête. Un modèle comprend donc la représentation des documents et des requêtes et définit une fonction de similarité permettant de déterminer la pertinence d'un document par rapport à une requête donnée. Ce score de pertinence entre un document d et une requête q est aussi appelé *Retrieval Status Value* (RSV) et souvent noté $RSV(d, q)$ (Nottelmann et Fuhr, 2003). Les modèles de RI peuvent être

classés en trois grandes catégories : les modèles booléens, les modèles vectoriels et les modèles probabilistes.

3.1 Le modèle booléen

C'est un modèle simple fondé sur la théorie des ensembles et l'algèbre de Boole (Baeza-Yates et Ribeiro-Neto, 1999). Dans ce modèle, chaque document est représenté par un ensemble de descripteurs. La requête, quant à elle, est exprimée sous forme d'une expression booléenne ; ses descripteurs sont reliés par des opérateurs booléens (AND, OR, et NOT) ; ce qui permet de formuler des requêtes expressives. Les poids associés aux descripteurs sont binaires, c'est-à-dire que si un descripteur est contenu dans un document (ou une requête), son poids pour ce document est 1, sinon son poids est nul. L'appariement entre un document et une requête est basé sur la satisfaction de l'expression booléenne. Autrement dit, seuls les documents qui satisfont exactement la requête sont retournés. C'est un modèle simple à implémenter et largement utilisé par des systèmes commerciaux mais qui présente les principales limites suivantes (Diallo, 2006):

- Avec sa pertinence binaire, un document est soit classé comme pertinent, soit comme non pertinent ; par conséquent, il n'est pas possible de classer les documents résultats. En plus, la pertinence binaire ne permet pas l'appariement approximatif entre documents et requêtes ;
- Tous les descripteurs appartenant à un document (ou une requête) ont le même poids même si certains sont en réalité plus pertinents que d'autres ;
- Il est difficile pour les utilisateurs d'exprimer leurs besoins en information par des expressions booléennes.

3.2 Le modèle vectoriel

Pour pallier les limites du modèle booléen et améliorer ainsi les performances des SRI, Salton et ses collègues ont proposé une extension du modèle booléen (Salton et al., 1983) consistant en la pondération des termes d'indexation des documents en tenant compte de leur importance relative dans la collection à traiter. Cette extension propose ainsi le schéma de pondération TF.IDF (term frequency – inverse document frequency), une mesure combinant le nombre d'occurrences du terme dans un document et sa fréquence documentaire dans la collection (cf section suivante). L'appariement documents-requête est basé sur un modèle appelé p-norm qui permet de répondre aux limites liées aux poids binaires mais aussi de prendre en compte l'ordonnancement des documents retrouvés.

Dans le modèle vectoriel (Vector Space Model ou VSM), les documents et les requêtes sont représentés par des vecteurs dans un espace multi-dimensionnel formé par les termes d'indexation (Salton et al., 1975; Salton et McGill, 1986). Les coordonnées d'un vecteur correspondent aux poids des termes d'indexation dans le document (ou la requête) qu'il représente. Pour déterminer le degré de correspondance entre un document et une requête, des mesures de similarité telles que le cosinus de l'angle formé par les vecteurs représentant ces derniers sont utilisées.

Considérons un document d et une requête q représentés dans un espace de dimension m respectivement par :

$$d = (w_1^d, w_2^d, \dots, w_m^d) \quad \text{et} \quad q = (w_1^q, w_2^q, \dots, w_m^q)$$

où w_i^d est le poids du terme t_i dans le document d et w_i^q le poids du terme t_i dans la requête q.

Le cosinus entre ces deux vecteurs, communément utilisé pour estimer la pertinence d'un document par rapport à une requête, est calculé comme suit :

$$RSV\ (d,q) = \frac{\sum_{i=1}^{m} w_i^d w_i^q}{\sqrt{\sum_i^m (w_i^d)^2}\sqrt{\sum_i^m (w_i^q)^2}}$$

D'autres mesures de similarité telles que la distance euclidienne , le coefficient de Jaccard (Jaccard, 1912) et le coefficient de corrélation de Pearson permettent aussi de calculer ce degré de correspondance.

Le modèle vectoriel est relativement flexible et, contrairement au modèle booléen, il permet un appariement partiel entre documents et requêtes; c'est-à-dire que le SRI peut retourner des documents qui répondent partiellement à une requête. Les documents retournés sont classés par ordre décroissant de pertinence par rapport à la requête. De plus, les schémas de pondération utilisés permettent de combiner différents facteurs pour mesurer les poids des termes d'indexation. La mesure TF.IDF, qui prend en compte la fréquence du terme dans le document (TF), la fréquence documentaire du terme (DF) voire la longueur du document, est un exemple de schéma de pondération communément utilisé dans ce modèle. Il existe plusieurs variantes pour le calcul du TF.IDF dont un exemple est :

$$TF_{t_i,d} = \sqrt{f_{t_i,d}} \quad , \quad IDF_{t_i} = 1 + \log\left(\frac{N}{n_i}\right) ,$$

$$w_i^d = \boldsymbol{TF.IDF_{t_i,d}} = TF_{t_i,d} \times IDF_{t_i}$$

où $f_{t_i,d}$ est le nombre d'occurrences du terme t_i dans le document d ;

N le nombre total de documents dans la collection;

et n_i le nombre de documents contenant le terme t_i.

Le facteur *TF* permet d'estimer la représentativité d'un terme d'indexation dans un document tandis que le facteur *IDF* permet de mesurer son importance pour discriminer les documents. Un terme très fréquent dans la collection est peu discriminant et donc probablement moins pertinent.

Une des principales limites du modèle vectoriel est qu'il ne prend pas en compte des associations existant potentiellement entre les termes d'indexation qui sont considérés dans ce modèle comme mutuellement indépendants. En effet, ces termes peuvent entretenir des relations qui incluent la synonymie, l'homonymie, la subsomption et l'association et ne pas

exploiter ce type d'information peut avoir des conséquences néfastes sur les performances d'un SRI. Ainsi, d'autres travaux ont proposé des modifications au modèle vectoriel pour surmonter cette limite. C'est dans ce sens que Wong et ses collègues ont proposé le modèle vectoriel généralisé (Generalized Vector Space Model - GVSM) prenant en compte des corrélations entre termes d'indexation (Wong et al., 1985). Dans le GVSM, la cooccurrence des termes d'indexation dans la collection est utilisée pour calculer la corrélation entre chaque paire de termes. Une approche similaire utilise la base lexicale WordNet (Fellbaum, 1998) pour déterminer d'éventuelles associations entre les termes d'indexation (Tsatsaronis et Panagiotopoulou, 2009) . L'analyse sémantique latente (LSA) (Deerwester et al., 1990a) est également une technique statistique communément utilisée pour traiter les problèmes liés à l'indépendance des termes. Ces deux dernières approches seront détaillées dans la section 5 consacrée à la présentation de travaux les mettant en œuvre.

Les modèles probabilistes complètent ces modèles classiques de RI. Nous en faisons un rapide survol dans la section qui suit.

3.3 Le modèle probabiliste

C'est un modèle fondé sur la théorie des probabilités. La correspondance entre un document et une requête est basée sur la probabilité que le document soit pertinent pour la requête (Rijsbergen, 1979; Robertson et al., 1981). Pour une requête q donnée, on suppose qu'il existe un ensemble de documents pertinents pour cette requête. Pour chaque document d, le modèle calcule la probabilité qu'il soit pertinent pour la requête q, soit $P(R|d,q)$ et la probabilité de non pertinence pour q, soit $P(\bar{R}|d,q)$. Le degré de pertinence entre le document d et la requête q est déterminé selon la formule suivante (Robertson et al., 1981):

$$RSV(d,q) = log\ \frac{P(R|d,q)}{P(\bar{R}|d,q)}$$

Cette formule de probabilité est simplifiée en utilisant le théorème de Bayes par :

$$RSV(d,q) = log\ \frac{P(d|R,q)}{P(d|\bar{R},q)}$$

Avec $P(d|R,q)$ et $P(d|\bar{R},q)$), les probabilités d'appartenance du document d respectivement à l'ensemble des documents pertinents pour q et à l'ensemble des documents non pertinents pour q.

La distribution du document d dans ces deux ensembles permet d'estimer sa pertinence par rapport à la requête q. Ce modèle est le *Binary Independence Model* (BIM) (Robertson et Jones, 1976).

Comme dans le modèle vectoriel, les documents trouvés sont classés dans l'ordre décroissant de leur score de pertinence par rapport à la requête. En plus, ce modèle supporte la RI itérative où l'utilisateur peut intervenir dans le processus pour améliorer les performances. Cependant, il reste sensible aux valeurs des paramètres (pour le calcul des probabilités) dont l'optimisation est coûteuse. Le modèle BM25 (Robertson et Walker, 1994; Spark Jones et al.,

2000) a été proposé pour pallier les défauts de ce modèle BIM et reste aujourd'hui l'un des modèles les plus performants observés en RI. Le système Okapi (Robertson et al., 1996) est un exemple d'implémentation de ce modèle.

A côté de ce modèle, il y a d'autres modèles basés sur la théorie des probabilités comme le modèle de langue qui a été proposé par (Ponte et Croft, 1998) et largement exploré (Hiemstra et al., 2004). Le principe de ce modèle est de déterminer les documents les plus susceptibles de générer la requête ; il repose sur l'hypothèse selon laquelle lorsqu'un utilisateur a un besoin en information, celui-ci a une idée assez précise des documents qui l'intéressent et qu'il formule sa requête pour retrouver ces derniers. Chaque document d est associé à un modèle de document noté M_d. Le score de pertinence entre le document d et une requête q est la probabilité que q puisse être générée par M_d. L'idée est donc d'estimer cette dernière probabilité plutôt que de calculer la similarité entre le document et la requête.

Dans la section 4, nous allons présenter les techniques utilisées pour évaluer les performances d'un SRI.

4 Evaluation d'un système de recherche d'information

En RI, l'évaluation est une étape essentielle pour estimer les performances d'une méthode ou d'un SRI. Elle permet de vérifier si un tel système donne les résultats attendus, le cas échéant, de déterminer les techniques qui affectent positivement ou négativement ses performances mais aussi de comparer les résultats de différents systèmes ou méthodes. Puisque l'évaluation dans une application réelle n'est pas très pratique, elle se réalise généralement sur des collections standard de tests. Une collection de tests est constituée comme suit ; pour chaque requête, on fournit une liste de documents avec un score de pertinence pour chacun de ces documents par rapport à la requête. Les documents non évalués sont considérés comme non pertinents. Les collections de tests sont souvent fournies dans des campagnes scientifiques d'évaluation organisées régulièrement dans un domaine donné (Harman, 1992; Cleverdon, 1991).

Pour l'évaluation, différentes mesures sont proposées. Dans cette section, nous présentons d'abord des mesures communément utilisées en RI avant de faire un aperçu des principales campagnes d'évaluation du domaine.

4.1 Les mesures d'évaluation d'un système de recherche d'information

Les mesures les plus couramment utilisées en RI sont la précision et le rappel. Ils permettent de mesurer la capacité du SRI à retrouver les documents pertinents et à rejeter ceux qui ne le sont pas. Un système idéal doit pouvoir retrouver seulement les documents pertinents vis-à-vis d'une requête. Pour une requête q donnée, les résultats peuvent être classés en deux catégories : ceux qui sont pertinents pour la requête et ceux qui ne le sont pas. Considérons DP, l'ensemble des documents pertinents pour q et, DR l'ensemble des documents retournés par le système. Les mesures de précision P et de rappel R sont définies respectivement par :

$$P = \frac{|DP \cap DR|}{|DR|}$$

$$R = \frac{|DP \cap DR|}{|DP|}$$

Autrement dit, la précision mesure la proportion de documents retrouvés qui sont pertinents tandis que le rappel est la proportion de documents pertinents qui sont retrouvés. Plus ces deux grandeurs sont importantes, plus les performances du système sont bonnes. Un SRI ayant une précision égale à 1 signifie que tous les documents qu'il a retournés sont pertinents. Il obtient un rappel égal à 1 s'il retrouve tous les documents pertinents. Toutefois, chacune de ces deux mesures permet d'estimer partiellement les performances d'un SRI et une évaluation basée sur les deux n'est pas simple. Par exemple, si un système A a une précision de 73% et un rappel de 45% et un système B une précision de 54% et un rappel de 65%, lequel de ces deux systèmes a une meilleure performance ? Répondre à cette question nécessite de trouver un compromis entre le rappel et la précision. Van Rijsbergen propose ainsi la f-measure qui permet de combiner ces deux mesures (Rijsbergen, 1979). Elle est définie comme suit :

$$F_\beta = \frac{(\beta^2 + 1)(P \times R)}{\beta^2 \times P + R}$$

où β est une constante qui détermine l'importance relative du rappel et de la précision.

Plus la f-mesure est élevée, plus la performance du SRI est bonne. Plus la valeur de β est importante, plus on privilégie la précision sur le rappel. Dans le cas particulier où β est fixé à 1, ces deux mesures sont d'égale importance et la f-mesure devient leur moyenne harmonique, notée $F1$ qui est communément utilisée dans la littérature:

$$F_1 = \frac{2 \times P \times R}{P + R}$$

Le pourcentage de documents pertinents sur les k meilleurs (par ordre de pertinence; k étant fixé) documents retournés par le SRI est aussi utilisé pour mesurer sa performance. En effet, dans la pratique, les utilisateurs sont généralement intéressés par les meilleurs résultats retournés par un SRI. Donc l'idée est d'estimer la précision du système en considérant seulement les k meilleurs documents qu'il a retournés, c'est à dire sa précision pour les k documents retrouvés en tête. Cette mesure courante est notée *P@k* et est définie formellement par :

$$P@k = \frac{\sum_{q \in Q} P(q,k)}{|Q|}$$

avec

$$P(q,k) = \frac{DP_k}{k}$$

et avec DP_k, le nombre de documents pertinents parmi les k premiers documents retournés pour la requête q.

Les précisions pour k = 5 (*P@5*) et k = 10 (*P@10*) sont communément utilisées dans les campagnes d'évaluation.

Une autre mesure similaire notée NDCG@k (Normalized Discounted Cumulated Gain) (dont la description complète peut être trouvée dans (Järvelin et Kekäläinen, 2002)) préconise l'utilisation d'une pertinence graduée (prise sur une échelle) plutôt que la pertinence binaire pour mesurer la performance d'un système. Les mesures de précision et de rappel supportent seulement les jugements de pertinence binaire. Par conséquent, elles ne permettent pas de distinguer les documents très pertinents des documents peu pertinents. Le NDCG est une mesure d'évaluation basée sur la pertinence graduée pour prendre en compte cet aspect. Cette mesure prend en considération le fait qu'il est préférable de présenter les documents très pertinents à la tête du classement par rapport à ceux qui sont peu pertinents. La pertinence est considérée ici comme une mesure d'utilité (ou de gain) pour un document à examiner. Un des inconvénients de cette métrique est que la production de jugements de pertinence graduée est plus difficile et plus coûteuse.

La précision moyenne (Average Precision – AP) est une autre mesure d'évaluation définie comme la moyenne arithmétique de la précision, pour une requête q, sur l'ensemble des positions des documents retournés :

$$AP(q) = \sum_{k=1}^{|DR|} \frac{P(q,k) \times rel(k)}{|DR \cap DP|}$$

avec $rel(k) = 1$ si le document se trouvant au rang k est pertinent, $rel(k) = 0$ sinon ; DP, l'ensemble des documents pertinents pour q et, DR l'ensemble des documents retournés.

Lorsque qu'un ensemble de requêtes Q est considéré, leur précision moyenne est définie par la MAP (Mean Average Precision) qui est une des mesures les plus couramment utilisées en RI :

$$MAP = \frac{\sum_{q \in Q} AP(q)}{|Q|}$$

Nous verrons dans le chapitre 5 qu'il existe d'autres mesures également utilisées pour évaluer les performances de systèmes de certains domaines spécifiques de la RI, comme la classification de documents à large échelle.

Dans la section suivante, nous présentons les campagnes d'évaluation au cours desquelles des méthodes et systèmes de RI sont confrontés.

4.2 Les campagnes d'évaluation

Il y a principalement deux campagnes internationales visant à évaluer les performances en RI : TREC et CLEF.

TREC[30] (Text REtrieval Conference) est un projet démarré en 1992 ayant pour but de fournir une infrastructure qui permet d'évaluer les méthodes et les systèmes de RI sur de grandes collections de tests. Le projet TREC consiste en une série d'ateliers (tracks) annuels dont l'objectif principal est d'encourager la recherche en RI mais aussi de fournir des techniques d'évaluation appropriées permettant aux chercheurs de mesurer les performances de leurs systèmes. Chaque atelier consiste en un ensemble de tâches focalisées sur un domaine de recherche particulier. Les ateliers évoluent et changent d'objectifs régulièrement en fonction de l'évolution des problématiques de recherche en RI. Ainsi, dans la campagne TREC 2014, huit ateliers ont été organisés et s'intéressaient à différentes questions telles que l'aide à la décision clinique (*Clinical Decision Support*), la suggestion contextuelle (*Contextual Suggestion*), la recherche dans les micro-blogs etc.

CLEF[31] (Conference and Labs of the Evaluation Forum connu formellement sous le nom Cross-Language Evaluation Forum) est une initiative démarrée en 2000 dont l'objectif est de promouvoir la recherche et l'innovation dans la RI. Elle se focalise sur les aspects multilingues et multimodaux en RI mais aussi sur l'utilisation des données non structurées, semi-structurées et structurées en RI. Ainsi, l'utilisation de ressources telles que les thésaurus et les ontologies est explorée. CLEF fournit un ensemble de collections de tests et supporte l'évaluation des SRI sur des données expérimentales. En plus d'une conférence internationale portant sur la recherche multilingue et les techniques d'évaluation, elle supporte une série d'ateliers (labs) sur la RI mono et translingue. Dans sa dernière campagne (CLEF 2014), huit ateliers ont été proposés dont *CLEF/ehealth Lab*[32] (Kelly et al., 2014) qui s'intéressait aux problématiques d'accès à l'information dans le domaine médical.

En plus de ces deux projets, on note aussi, plus récemment, la campagne BioASQ[33] (Tsatsaronis et al., 2012) issue du projet européen du même nom, démarré en 2013, qui s'intéresse à des aspects particuliers de la RI dans le domaine biomédical. Actuellement, elle comprend deux tâches : la *tâche A* qui concerne l'indexation sémantique à grande échelle de documents biomédicaux et la *tâche B* qui s'intéresse aux systèmes de Questions/Réponses (QA) dans le domaine biomédical. Ce challenge dispose de larges collections de tests et une infrastructure d'évaluation en ligne qui permet de tester facilement des méthodes ou systèmes. Nous évoquerons plus en détail cette campagne dans le chapitre 5 car nous avons utilisé ces collections pour l'évaluation de notre approche de classification.

Ces différentes campagnes d'évaluation ont contribué au développement de modèles de RI plus performants qui sont couramment utilisés dans des applications pratiques. Toutefois, la plupart des systèmes classiques de RI sont basés sur un modèle à sacs de mots clés et restent ainsi confrontés à des problèmes liés au traitement de la langue naturelle, tels que l'ambiguïté des mots et la disparité des termes (termes lexicalement différents mais ayant des significations proches). C'est ce qui a motivé le développement des nouvelles approches de

[30] http://trec.nist.gov/
[31] http://www.clef-initiative.eu/
[32] http://clefehealth2014.dcu.ie/
[33] http://www.bioasq.org/

RI s'appuyant sur la sémantique véhiculée par le contenu des documents que nous présentons dans la section suivante.

5 La recherche sémantique d'information

Comme présenté précédemment, les méthodes classiques sont généralement basées sur la représentation par mots clés, appelée aussi la représentation par *sac-de-mots*; les documents et les requêtes sont décrits par un ensemble de mots (ou même des n-grammes) qu'ils contiennent; la correspondance entre un document et une requête est basée sur le nombre de mots qu'ils partagent. En plus de ne pas tenir compte de l'ordre des mots (*communication technique* vs *technique de communication*) et de se baser sur un appariement purement lexical (matching 1:1 entre mots), elles sont ainsi affectées par deux limites principales : l'ambiguïté (polysémie et homonymie) des mots et la synonymie. Un terme peut être ambigu (avoir plusieurs sens) et sa signification dépend dans ce cas de son contexte d'utilisation. Par exemple, le terme *diabetes* est associé à trois concepts différents dans l'UMLS. Lorsqu'il y a des termes ambigus parmi les descripteurs, le système peut retourner des documents contenant des termes de la requête qui ne sont pourtant pas pertinents (baisse de la précision). Un terme peut aussi avoir plusieurs synonymes. Par exemple, les termes *cancer* et *malignancy* sont des synonymes dans l'UMLS. Dans ce cas, si on effectue une recherche en utilisant un de ces synonymes, les documents pertinents mais décrits par les autres synonymes peuvent ne pas être retrouvés (baisse du rappel). La RI classique considère généralement les termes d'indexation comme des entités indépendantes et par conséquent ne permet pas la prise en compte des associations entre elles. Intuitivement, on se doute que les relations sémantiques (synonymie, méronymie, hyponymie, etc) pourraient être bénéfiques pour améliorer les performances d'un SRI. Par exemple, un utilisateur qui cherche des informations sur la *démence* peut être intéressé par des documents sur la *maladie d'Alzheimer*, qui est une sorte de *démence*. Nous rappelons que les thésaurus sont aussi largement explorés pour ces tâches.

La recherche sémantique, définie comme la recherche basée sur la sémantique des termes, a été proposée pour surmonter ces problèmes et améliorer ainsi les performances des SRI. L'idée est de prendre en compte le contenu sémantique véhiculé par les documents et les requêtes plutôt que de les décrire par de simples sacs de mots. Vu l'intérêt qu'elle a suscité, la RI sémantique a fait l'objet de nombreux travaux de recherche ces dernières années.

Dans cette section, nous passons en revue les principales approches qui ont été proposées et qui peuvent être classées en deux grandes catégories : 1) les approches statistiques généralement basées sur la sémantique vectorielle qui exploite l'information intrinsèque contenue dans les documents et 2) les approches utilisant des ressources sémantiques externes, telles que les thésaurus ou les ontologies.

5.1 Les approches statistiques

Une des techniques statistiques les plus couramment utilisées en RI pour pallier les limites liées à l'ambiguïté et la synonymie des termes est la LSA (Deerwester et al., 1990a). La LSA utilise des techniques algébriques basées sur la fréquence et la cooccurrence des mots dans les

documents pour construire des « concepts ». A partir d'une matrice termes-documents décrivant les occurrences des termes d'indexation dans les documents, elle utilise la décomposition en valeurs singulières (Singular Value Decomposition - SVD), une technique de réduction de dimension similaire à l'analyse factorielle, pour produire un espace conceptuel d'indexation de taille réduite (Dumais, 1994). Dans ce nouvel espace, les termes sont regroupés sous forme de « concepts » où un « concept » est représenté par un vecteur pondéré de mots sémantiquement proches. Ces derniers permettent ainsi de capturer la sémantique latente des documents. Les « concepts » résultants, qui constituent les nouvelles dimensions de l'espace d'indexation, permettent une représentation sémantique des documents. Ainsi, les documents partageant des termes co-occurrents (donc des concepts), sont considérés comme sémantiquement proches et ont donc des représentions vectorielles proches. De même, des termes avec des expressions différentes peuvent être sémantiquement proches. L'application de la LSA en RI permet ainsi un appariement document-requête plus flexible basé sur la sémantique latente plutôt que sur les mots qu'ils partagent. Ainsi, un document peut avoir un score de pertinence élevé par rapport à une requête sans partager pour autant les mêmes mots.

L'avantage de cette approche est qu'elle ne nécessite pas de connaissances externes. Par conséquent, elle reste indépendante et applicable dans tout domaine. De plus, les associations entre les termes (synonymie, polysémie) sont partiellement prises en compte. Dans de nombreuses expérimentations, elle a obtenu des performances meilleures que les techniques classiques de RI (Dumais, 1994; Letsche et Berry, 1997) mais reste limitée pour le traitement de grandes collections de documents (Letsche et Berry, 1997) puisqu'elle dépend du SVD dont le calcul est lourd. De plus, la mise à jour des index est difficile lorsque de nouveaux documents sont intégrés.

Des méthodes similaires, basées sur les techniques de réduction de dimension, ont été proposées dans la littérature. Huang et ses collègues proposent ainsi une approche combinant les contextes locaux et globaux des mots pour capturer leur sémantique dans les documents (Huang et al., 2012). Le contexte local d'un mot est constitué des mots l'entourant tandis que son contexte global est le document entier le contenant. Mikolov et ses collègues, quant à eux, proposent une représentation vectorielle des mots du corpus dans un espace continu de dimension plus petite (Mikolov et al., 2013). Pour cela, ils s'appuient sur les réseaux neuronaux. Ainsi, les mots sont représentés par des vecteurs qui capturent leur sémantique. Ces représentations permettent ainsi de capturer les associations entre les différents termes plutôt que de les considérer comme indépendants, ce qui est très intéressant en RI.

Bien que ces approches algébriques et statistiques permettent d'atteindre de bonnes performances en RI, l'utilisation de connaissances explicites semble être nécessaire pour améliorer significativement les performances des SRI (Croft, 1986). Ainsi, des approches alternatives proposent d'exploiter les connaissances explicites contenues dans des ressources sémantiques afin d'améliorer les performances des SRI.

5.2 Les approches basées sur des ressources externes

Face aux limites des méthodes classiques de RI, dues principalement aux problèmes de représentation et d'appariement lexicaux, d'autres approches s'appuient sur des ressources sémantiques telles que les thésaurus ou les ontologies. Actuellement, les ressources sémantiques sont de plus en plus utilisées en RI pour supporter soit l'indexation des documents par leur contenu sémantique, soit l'expression précise du besoin en information ou encore la reformulation et l'expansion des requêtes mais également l'appariement des documents et des requêtes. Cet engouement est illustré par les nombreux travaux dans le domaine durant la dernière décennie et qui ont abouti à des résultats intéressants.

Dans cette section, nous présentons d'abord des ressources couramment utilisées en RI sémantique. Ensuite, les mesures de similarité sémantique et les différentes approches basées sur ces ressources sont exposées.

5.2.1 WordNet

WordNet (Fellbaum, 1998) est une des ressources les plus couramment utilisées par les modèles de RI sémantique. C'est une ressource linguistique riche qui couvre divers domaines. Elle couvre largement la langue anglaise (noms, verbes, adjectifs et adverbes) et contient, en plus de ces informations linguistiques, des connaissances ontologiques. Les termes sont organisés en des ensembles de synonymes, appelés synsets (qui représentent des concepts). Chaque synset représente un sens spécifique d'un terme ; chaque terme peut désigner un mot simple ou un groupe de mots. En général, chaque terme est associé à un ou plusieurs synsets (polysémie). Les synsets sont structurés par plusieurs types de relations sémantiques (hyperonymie, hyponymie, méronymie, etc.). Dans sa version 3 actuelle, ses 155 287 termes sont regroupés en 117 659 synsets. En plus de sa large couverture et de sa richesse, WordNet[34] est librement accessible, facilitant d'autant plus son exploitation en RI.

Au-delà des ressources générales telles que WordNet, DBpedia[35], Wikipédia[36], YAGO[37], des ressources spécifiques, déjà décrites dans le chapitre précédent, telles que l'UMLS, le thésaurus MeSH, la Gene Ontology (GO) sont couramment utilisées dans le domaine médical.

5.2.2 Les mesures de similarité sémantique

Une mesure de similarité sémantique permet d'estimer à quel point deux concepts (ou deux termes) sont sémantiquement proches. Dans la littérature, de nombreuses techniques sont proposées pour calculer cette similarité en utilisant une ressource sémantique. La plupart de ces techniques s'appuient sur les relations taxonomiques.

Nous présentons dans cette section les mesures de similarité les plus courantes que l'on peut classer en deux catégories: les mesures reposant sur la longueur du chemin liant deux concepts et celles utilisant leur contenu en information.

[34] http://wordnet.princeton.edu/
[35] dbpedia.org
[36] https://www.wikipedia.org/
[37] http://www.mpi-inf.mpg.de/yago-naga/yago/

5.2.2.1 Les mesures de similarité basée sur la distance entre deux concepts

Dans cette catégorie de mesures, la similarité sémantique entre deux concepts dépend du nombre de nœuds (concepts) dans le chemin le plus court séparant ces deux concepts et de leur(s) position(s) dans la hiérarchie.

Ainsi, la mesure proposée dans (Rada et al., 1989) considère la similarité sémantique entre deux concepts comme l'inverse de la longueur du plus court chemin séparant les deux concepts. Leacock et Chodorow définissent une mesure de similarité comparable où la profondeur de la taxinomie est considérée (Leacock et Chodorow, 1998) :

$$Sim_{Leacock}(C_1, C_2) = -\log\left(\frac{\mathrm{p}}{2 * \mathrm{depth}}\right)$$

où p est le nombre de nœuds dans le chemin le plus court entre C_1 et C_2 et $depth$ la profondeur maximale de la taxonomie.

Une autre métrique proposée dans (Wu et Palmer, 1994) considère la similarité sémantique entre deux concepts comme une fonction de leurs profondeurs dans la taxonomie mais aussi de celle du concept le plus spécifique qui les subsume :

$$Sim_{Wu}(C_1, C_2) = \frac{2 * \mathrm{depth}(C)}{\mathrm{depth}(C_1) + \mathrm{depth}(C_2)})$$

avec C le concept le plus spécifique qui subsume C_1 et C_2, $depth(C_i)$, la profondeur du concept C_i dans la taxonomie.

En plus de ces mesures les plus courantes, une variété de grandeurs pour le calcul de similarité entre concepts est proposée dans la littérature. Pour plus de détails, nous renvoyons le lecteur à la revue faite dans (Sánchez et Batet, 2011).

A côté de ces métriques basées seulement sur la distance entre les concepts et leurs profondeurs, d'autres travaux proposent d'exploiter le contenu informationnel des concepts.

5.2.2.2 Les mesures de similarité basée sur le contenu informationnel

Le contenu informationnel d'un concept est une mesure visant à estimer sa spécificité. Il est généralement déterminé en fonction de la probabilité d'occurrence du concept dans un corpus préalablement défini et en rapport avec la ressource sémantique utilisée. On considère que les concepts généraux véhiculent moins de contenu informationnel que les concepts spécifiques.

Pour estimer le degré de similarité entre deux concepts, Resnik propose une métrique basée sur le contenu informationnel du concept le plus spécifique qui les subsume (« least common subsumer ») (Resnik, 1999). Cette mesure considère que deux concepts sont sémantiquement proches si la quantité d'information qu'ils partagent (le contenu informationnel de leur parent commun) est importante. Ainsi, la similarité sémantique pour une paire de concepts C_1 et C_2 est définie par :

$$Sim_{Resnik}(C_1, C_2) = IC(C)$$

avec C le concept le plus spécifique qui subsume C_1 et C_2, $IC(C)$, le contenu informationnel du concept C.

Dans la littérature, il existe plusieurs variantes pour estimer le contenu informationnel d'un concept. Dans (Resnik, 1999), il est défini par :

$$IC(C) = -\log\left(\frac{freq\ (C)}{freq\ (root)}\right)$$

avec $freq\ (C)$ (respectivement $freq\ (root)$), la somme de la fréquence d'occurrence du concept C (respectivement $root$, le concept noyau) plus celles de ses hyponymes dans le corpus.

Dans (Jiang et Conrath, 1997), les auteurs proposent une mesure combinant les contenus informationnels des concepts et celui de leur parent commun. Cette mesure est définie par la formule suivante :

$$Sim_{Jiang}(C_1, C_2) = \frac{1}{IC(C_1) + IC(C_2) - 2 * IC(C)}$$

avec C et $IC(C)$ identiques que dans la formule de sim_{Resnik}.

Une mesure similaire utilisant les mêmes informations est proposée dans (Lin, 1998) :

$$Sim_{Lin}(C_1, C_2) = \frac{2\ *\ IC(C)}{IC(C_1)\ +\ IC(C_2)}$$

avec les mêmes définitions de C et $IC(C)$.

A l'instar de celles basées sur les chemins entre concepts, une variété de mesures de similarité utilisant le contenu informationnel a également été développée. Pour une revue complète de ces différentes travaux, le lecteur peut se référer à (Sánchez et Batet, 2011).

Plusieurs travaux se sont intéressés à l'évaluation et à la comparaison de ces différentes mesures de similarité (Hliaoutakis et al., 2006; Pedersen et al., 2007; Batet et al., 2011; Garla et Brandt, 2012). Dans leurs expérimentations, Hliaoutakis et ses collègues (2006) ont montré que les mesures de similarité sémantique utilisant le contenu informationnel donnent les meilleurs résultats sur WordNet tandis que sur MeSH, les deux approches donnent des résultats comparables. Garla et Brandt (2012) ont évalué différentes mesures de similarité sur des benchmarks en utilisant différentes ressources sémantiques médicales (SNOMED CT, MeSH, et l'UMLS). A travers cette évaluation, les auteurs ont démontré que le calcul de similarité en utilisant l'UMLS entièrement donne des résultats meilleurs qu'en se basant seulement sur la SNOMED CT ou le thésaurus MeSH. Là encore, les mesures basées sur le contenu informationnel se sont révélées largement plus performantes. Pederson et ses collègues ont aussi expérimenté ces mesures en utilisant la SNOMED-CT (Pedersen et al., 2007) et l'UMLS (McInnes et Pedersen, 2013). Ils ont eux aussi noté que les mesures basées sur le contenu informationnel résultent en des performances plus élevées que les mesures utilisant des chemins pour une tâche de désambiguïsation (McInnes et Pedersen, 2013).

5.2.3 Les approches de recherche d'information basées sur des ressources sémantiques externes

Pour pallier les limites des approches classiques de RI, différentes méthodes d'indexation sémantique ont été mises en œuvre. Pour résoudre les problèmes liés à l'ambigüité des termes, des techniques de désambiguïsation (Word Sense Disambiguation ou WSD), généralement basées sur des ressources sémantiques externes et les contextes d'utilisation des mots, ont été proposées. Le principe est de retrouver, pour chaque terme d'indexation, ses différents sens (appelés dans certains cas « concepts ») à partir d'une ressource externe, et d'exploiter son contexte local (concepts proches tels que le(s) parent(s) et le(s) enfant(s)) pour déterminer le sens concerné. La disparité des termes est, quant à elle, souvent traitée en utilisant des techniques d'expansion de requêtes (Hersh et al., 2000). L'indexation conceptuelle où les documents (et les requêtes) sont représentés par des concepts d'un thésaurus ou d'une ontologie, a également émergé pour surmonter ces limites.

Puisque le développement de ces modèles de connaissances est coûteuse, des ressources existantes générales, telles que WordNet, ou plus spécifiques à un domaine particulier, telles que MeSH en médecine, ont été largement exploitées en RI sémantique. Elles ont été utilisées pour traiter différentes questions : la désambiguïsation de termes décrivant les requêtes et les documents (Voorhees, 1993), l'expansion de requêtes avec des termes sémantiquement proches (Voorhees, 1994), la représentation et la comparaison des requêtes et des documents dans un espace conceptuel (Gonzalo et al., 1998) ou encore l'identification d'associations entre les termes (Tsatsaronis et Panagiotopoulou, 2009).

5.2.3.1 Utilisation de ressources linguistiques

Désambiguïsation des termes

Voorhees a proposé une approche d'indexation basée sur le sens des mots (Voorhees, 1993). Ces derniers sont extraits et projetés dans WordNet pour récupérer les synsets correspondants. Lorsqu'un mot correspond à plusieurs synsets (ambigu), son contexte local est exploré pour déterminer le sens approprié. Pour cela, l'auteur considère les mots communs entre le voisinage de chaque synset dans WordNet et son contexte local. Toutefois, les expérimentations montrent que cette approche obtient des performances inférieures comparativement à une approche classique utilisant simplement les mots clés.

Dans un travail similaire (Katz et al., 1998; Uzuner et al., 1999), les auteurs ont développé une méthode de désambiguïsation de mots basée sur leurs contextes locaux. Les mots sont extraits et projetés dans WordNet, comme dans l'approche précédente, pour identifier leurs synsets associés. Dans le cas où le mot est ambigu, son contexte local, constitué des mots de son voisinage dans le document, est exploité pour identifier le sens adéquat. Les auteurs considèrent que les mots utilisés dans le même contexte sont souvent sémantiquement proches. Cette hypothèse est similaire à celles utilisées dans les techniques statistiques généralement basées sur la cooccurrence des mots pour capturer leur sémantique (Deerwester et al., 1990a). Pour chaque mot, les mots se trouvant dans son contexte sont extraits et projetés dans WordNet. Parmi les synsets qui lui sont associés, celui qui partage le plus de

mots avec l'ensemble des synsets de son contexte est retenu comme correspondant au sens approprié. L'évaluation de cette méthode de désambiguïsation sur le corpus Semcor a donné une précision de 60 % mais son incorporation dans le système SMART (Buckley et al., 1995) n'a pas permis, comme espéré, d'améliorer les performances du SRI. Les auteurs considèrent que ces résultats sont dus au fort taux d'erreurs de leur méthode de désambiguïsation.

Gonzalo et ses collègues (Gonzalo et al., 1998) se sont également intéressés à l'impact de la désambiguïsation sur les performances de leur SRI. Des expérimentations sur une collection de textes (requêtes et documents) désambiguïsée manuellement ont montré une amélioration considérable des performances en RI (+ 29 % par rapport aux méthodes classiques). Ils ont évalué la sensibilité des erreurs de désambiguïsation sur les résultats et ont conclu que jusqu'à un taux d'erreurs de 30 %, leur méthode utilisant les synsets de WordNet donne des performances meilleures qu'une méthode classique basée sur les mots clés. Une des limites de cette approche reste la désambiguïsation manuelle des documents et des requêtes.

Guarino et ses collègues ont implémenté une approche de RI guidée par l'ontologie Sensus, une large ontologie linguistique basée sur WordNet (Guarino et al., 1999). Les ressources et les requêtes sont décrites dans un graphe conceptuel simple et expressif où chaque élément correspond à un nœud de l'ontologie. En cas de polysémie, une désambiguïsation manuelle est réalisée pour sélectionner le sens adéquat. Ils ont développé un système dénommé OntoSeek (Guarino et al., 1999) basé sur cette méthode pour la recherche en ligne des pages jaunes et de catalogues de produits. Les auteurs ont montré le potentiel des ontologies qui améliorent significativement les performances de leur SRI. Baziz et ses collègues ont développé une méthode d'indexation sémantique qui inclut une phase de désambiguïsation des termes (Baziz et al., 2005). L'approche de désambiguïsation proposée s'appuie sur les mesures de similarité sémantique. Elle est fondée sur l'hypothèse suivante : si un terme est polysémique, son sens le plus adéquat est celui qui est le plus sémantiquement proche des autres termes contenus dans son contexte (termes co-occurrant avec un terme dans une fenêtre d'un nombre de mots donné). Pour désambiguïser un terme ambigu, ses synsets correspondants (qui sont les concepts candidats) sont récupérés de même que les synsets associés aux termes de son contexte. Pour chaque concept candidat, son degré (score) de similarité par rapport aux autres concepts du contexte est calculé et le candidat possédant le score maximal est retenu. Une fois les concepts désambiguïsés, ils sont utilisés pour représenter les documents et les requêtes dans un réseau conceptuel. Dans ce réseau, les liens entre concepts sont pondérés par les valeurs de leur similarité sémantique.

Expansion de requêtes

L'approche proposée dans (Voorhees, 1994) utilise WordNet pour effectuer une expansion manuelle des requêtes en utilisant les relations de synonymie, d'hyperonymie et d'hyponymie. Les résultats des expérimentations ont montré une amélioration des performances sur des requêtes courtes. Mais pour des requêtes longues, aucune amélioration significative n'a été observée par rapport aux approches de RI classiques.

Dans (Gonzalo et al., 1998), l'utilisation de WordNet a permis d'améliorer les résultats de la recherche en appliquant le modèle vectoriel sur des synsets plutôt que sur des mots clés. Les

résultats de leurs expérimentations montrent une amélioration du rappel qui atteint 62 % avec une indexation par les synsets contre 53,2 % pour une indexation par le sens des termes (les différents sens des mots sont distingués) et 48 % pour une méthode d'indexation classique (en utilisant le système de RI SMART).

Des approches similaires ont été proposées dans (Mihalcea et Moldovan, 2000; Sanderson, 1994; Liu et al., 2004). La méthode présentée dans (Liu et al., 2004) a donné des résultats intéressants avec une amélioration des performances de 23 à 30 % sur les collections TREC 9, 10 et 12 en utilisant des requêtes courtes.

Au-delà de la désambiguïsation et de l'expansion de requêtes, d'autres travaux ont exploité WordNet pour améliorer l'appariement des documents aux requêtes.

Appariement document – requête

Le modèle proposé dans (Tsatsaronis et Panagiotopoulou, 2009), nommément le GVSM (Generalized Vector Space Model), étend le modèle vectoriel et exploite la similarité sémantique entre les termes dans l'appariement document-requête. Pour cela, il se base sur WordNet. Les auteurs soutiennent que la prise en compte des liens sémantiques entre les concepts représentant les documents et ceux des requêtes est une bonne stratégie pour améliorer les performances d'un modèle de RI (VSM). Ils proposent ainsi une extension du modèle vectoriel en tenant compte de la proximité sémantique entre les concepts. L'évaluation de cette méthode sur trois collections de tests (TREC 1, 4 et 6) a permis de confirmer leur hypothèse.

L'utilisation de WordNet en RI a ainsi été investiguée dans de nombreux travaux. Dans certaines approches, elle a permis d'accroitre significativement les performances (Gonzalo et al., 1998; Liu et al., 2004), tandis que dans d'autres, les performances sont moins bonnes ou même se dégradent (Voorhees, 1993; Katz et al., 1998).

En plus de WordNet, la large base de connaissances Wikipédia a été également explorée plus récemment. L'analyse sémantique explicite (explicit semantic analysis – ESA), qui repose sur Wikipédia, est une approche proposée pour la représentation sémantique des documents textuels (Gabrilovich et Markovitch, 2006; Gabrilovich et Markovitch, 2007). Les documents sont représentés dans un espace conceptuel de grande dimension constitué de concepts extraits automatiquement à partir de Wikipédia. Cette méthode été appliquée avec beaucoup de succès en RI avec une amélioration conséquente des performances (Egozi et al., 2011).

5.2.3.2 Utilisation de ressources sémantiques spécifiques

Au-delà des ressources généralistes telles que WordNet, d'autres travaux ont utilisé des ressources sémantiques plus spécifiques pour accroître les performances en RI. Avec le développement du Web sémantique, des chercheurs ont également exploré le potentiel des ontologies dans ce domaine.

Khan et al. ont proposé un modèle d'indexation conceptuelle basée sur une ontologie de domaine du sport (Khan et al., 2004). Les auteurs adressent la question de l'identification des

concepts associés aux termes repérés dans les documents. Ils considèrent que, même dans une ontologie, un terme peut être associé à plusieurs concepts et proposent ainsi un algorithme de désambiguïsation basé sur la cooccurrence et la proximité sémantique pour éliminer les concepts non pertinents. Ils ont également défini un mécanisme d'expansion de requêtes contrôlée qui garantit la non-dégradation de la précision. Les résultats des expérimentations montrent une amélioration significative des performances de ce modèle comparativement à une méthode d'indexation classique : 91.3 % vs 55.6 % pour le rappel et 87.7 % vs 67.5 % pour la précision.

Plus récemment, des travaux tenant compte des liens éventuels entre concepts se sont attachés à développer des modèles d'indexation plus avancés. Ainsi, en se basant sur le modèle vectoriel généralisé (GVSM) (Wong et al., 1985), Hliaoutakis et ses collègues ont proposé un schéma de pondération où les liens entre concepts sont matérialisés par leur similarité sémantique (Hliaoutakis et al., 2006). Ils s'appuient sur le modèle vectoriel et proposent une pondération qui favorise les concepts fortement proches sémantiquement des autres concepts du document. Le modèle proposé, nommé SSRM (Semantic Similarity based Retrieval Model), repose sur la similarité sémantique entre les concepts d'une ressource sémantique. Le schéma TF.IDF est utilisé pour déterminer le poids initial d'un concept dans un document (ou une requête). Ce poids est ensuite recalculé pour les concepts de la requête en tenant compte de leurs degrés de similarité sémantique les uns par rapport aux autres. En plus des concepts contenus dans la requête, ceux qui sont sémantiquement similaires à ces derniers sont aussi utilisés pour l'étendre. Enfin, la correspondance entre un document et une requête est calculée en associant les concepts du document et ceux de la requête qui sont similaires. Cette fonction de similarité est définie, pour un document d et une requête étendue q, par :

$$Sim(q,d) = \frac{\sum_i \sum_j q_i d_j \, sim(c_i, c_j)}{\sum_i \sum_j q_i d_j}$$

où c_i et c_j correspondent respectivement aux concepts de la requête et du document.

L'évaluation de cette méthode sur la collection OHSUMED (une collection TREC de 293 856 articles médicaux), en utilisant les descripteurs du thésaurus MeSH, montre une amélioration significative des performances par rapport à une méthode de RI classique (VSM). Sa comparaison aux méthodes de RI sémantique telle que celle présentée dans (Voorhees, 1994) confirme également ses bonnes performances.

A côté de ces techniques utilisant une simple indexation conceptuelle, d'autres travaux ont exploré davantage le potentiel des ontologies et se sont intéressés à l'exploitation des bases de connaissances qui leur sont associées (Kiryakov et al., 2004; Bhagdev et al., 2008; Fernández et al., 2011). Au-delà des concepts, ces méthodes se focalisent sur les connaissances qu'ils représentent. Ainsi, les relations entre les entités de l'ontologie et les documents sont établies via des annotations. Pour cela, des techniques d'annotation sémantique (semi-automatique) sont exploitées ainsi que des langages formels comme SPARQL pour l'expression des requêtes. Cependant, elles ont été confrontées à quelques difficultés : l'utilisabilité des langages formels qui sont complexes, la question de l'ordonnancement des résultats, le

problème de la couverture des ontologies. Ainsi, des propositions ont été faites pour traiter le classement des résultats en adaptant le modèle vectoriel dans une approche de RI basée sur une ontologie (plus une base de connaissances) (Castells et al., 2007). Une combinaison de la RI sémantique et de la RI par mots clés a été également proposée pour faire face aux problèmes soulevés par l'incomplétude des ontologies (Bhagdev et al., 2008; Fernández et al., 2011).

En résumé, dans la plupart des approches proposées, les documents et les requêtes sont représentés par des ensembles de concepts. Dans cette représentation, analogue à la RI classique, où les concepts d'une ressource sémantique constituent l'espace d'indexation, un document (ou une requête) est généralement considéré comme un « sac de concepts ». Ensuite, des modèles classiques tels que le modèle vectoriel sont appliqués aux concepts (Gonzalo et al., 1998; Baziz et al., 2005). Dans cette représentation, la prise en compte des associations entre concepts fait défaut car ils sont pris indépendamment les uns des autres. Ceci peut avoir des effets néfastes sur la performance d'un SRI, comme nous l'avons déjà souligné précédemment (Section 3.2). Toutefois, dans l'indexation conceptuelle, les techniques d'expansion de requêtes sont souvent utilisées pour surmonter cette limite. D'autres travaux exploitent la similarité sémantique dans l'appariement des documents aux requêtes (Tsatsaronis et Panagiotopoulou, 2009; Hliaoutakis et al., 2006).

Dans la prochaine section, nous présentons plus spécifiquement les travaux sur la RI sémantique dans le domaine biomédical.

6 Recherche d'information sémantique dans le domaine biomédical

Dans le domaine médical, où le volume de données augmente de plus en plus, des chercheurs se sont également intéressés à la problématique de l'accès à l'information, et notamment aux problèmes soulevés par les méthodes de RI classique. De plus, l'abondance de ressources sémantiques, bien structurées (GO, SNOMED) et très vastes (UMLS), a motivé le développement de la RI guidée par ces dernières. Ainsi si certains travaux ont porté sur la désambiguïsation des termes médicaux (Widdows et al., 2003; Garla et Brandt, 2012; McInnes et Pedersen, 2013), d'autres se sont intéressés à l'identification des concepts dans des textes mais aussi à la sélection de concepts pertinents pour représenter un document.

6.1 Désambiguïsation des termes

Pour la désambiguïsation, à l'instar des méthodes générales présentées dans la section précédente, celles proposées dans le domaine biomédical exploitent le contexte local du concept (associé au terme). Le sens approprié d'un concept ambigu est déterminé en fonction de sa proximité par rapport aux autres concepts de son contexte. Pour estimer cette proximité, certaines méthodes s'appuient sur des relations sémantiques extraites à partir d'une source de connaissances (Widdows et al., 2003) tandis que d'autres reposent sur des techniques de calcul de similarité plus avancées (Dinh et Tamine, 2010; Garla et Brandt, 2012; McInnes et Pedersen, 2013).

Ainsi, pour un concept ambigu, ses différents sens sont retrouvés et pour chaque sens, son score de proximité est calculé. Ensuite, le sens ayant le score maximal est retenu.

Dans la section suivante, nous présentons tout d'abord différentes approches d'extraction de concepts à partir de textes médicaux (6.2). Ensuite, nous abordons les techniques d'indexation conceptuelle (6.3) avant de terminer par l'expansion de requêtes (6.4).

6.2 L'extraction de concepts médicaux

Concernant l'indexation conceptuelle, l'identification des concepts médicaux à partir de documents textuels est une étape préalable et primordiale. Elle peut être vue comme un sous-problème de la reconnaissance d'entités nommées (*named entity recognition* ou simplement *NER*) (Zhang et Elhadad, 2013). Cette dernière est une tâche très importante de l'extraction d'information. Habituellement, la reconnaissance d'entités nommées consiste à identifier et à classifier des éléments d'un texte dans des catégories prédéfinies, telles que des noms de personnes, d'organisations, de lieux, des dates, etc. (Nadeau et Sekine, 2007). Elle peut concerner aussi des catégories plus fines (protéines, gènes, etc.) dans certains contextes. On distingue principalement quatre approches pour repérer ces entités dans des corpus de textes : 1) les approches à base de dictionnaires (Zhou et al., 2006a; Zhang et Elhadad, 2013), 2) les approches basées sur des règles linguistiques (Subramaniam et al., 2003), 3) les approches statistiques généralement basées sur des techniques d'apprentissage automatique (He et Kayaalp, 2008; Minard et al., 2011) et 4) les approches hybrides qui combinent certaines de ces différentes techniques. Dans le domaine biomédical où les RTO sont particulièrement nombreuses, des dictionnaires (appelés aussi *gazetteers*), listant les entités cibles, sont habituellement utilisés (Bodenreider, 2006; Deléger et al., 2010). Cette tâche de reconnaissance d'entités médicales consiste à identifier les termes dénotant des concepts biomédicaux dans des corpus et à les classifier dans des catégories sémantiques définies dans une ressource spécifique. Toutefois, cette tâche est complexe vu la variation terminologique, la forte présence d'acronymes et d'abréviations mais aussi l'évolution rapide de la terminologie (nouveaux termes) dans le domaine médical.

Dans cette section, nous présentons ces différentes approches.

6.2.1 Les approches à base de dictionnaires

Les approches à base de dictionnaires utilisent des ressources sémantiques (thésaurus, ontologie, etc.) qui couvrent l'ensemble des concepts d'intérêt. Un exemple de cette catégorie est la méthode implémentée dans MetaMap (Aronson, 2001), un outil développé par la NLM (National Library of Medicine) pour l'identification de termes dénotant des concepts de l'UMLS à partir de textes. MetaMap réalise une analyse syntaxique pour extraire les syntagmes à partir du texte. Ensuite, les variantes de ces syntagmes sont générées en exploitant les connaissances contenues dans le lexique spécialiste[38] et alignées aux concepts du Metathesaurus de l'UMLS. Les concepts trouvés sont associés à des scores de confiance indiquant leur degré de correspondance au texte. La figure 6 montre un exemple de sortie de

[38] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

MetaMap pour la phrase *Dementia with Lewy bodies and Parkinson's disease dementia*. Des évolutions de MetaMap ont permis d'intégrer un module de désambiguïsation de concepts et de détecter la négation (Aronson et Lang, 2010). C'est l'un des outils les plus couramment utilisés pour la reconnaissance des concepts biomédicaux dans des corpus de textes. En plus de la littérature biomédicale, MetaMap a été largement exploité pour l'extraction d'information dans des textes cliniques (Schadow et McDonald, 2003; Meystre et Haug, 2006; Meystre et al., 2008).

Dans (Zhou et al., 2006a), une autre méthode à base de dictionnaires est proposée et implémentée dans un outil nommé MaxMatcher. Pour traiter les problèmes liés à la variation de la terminologie biomédicale, les auteurs proposent une méthode de correspondance approximative à la place d'un alignement exact. Ainsi, pour un concept donné, ils considèrent seulement les mots significatifs plutôt que tous les mots associés à ce concept. Le problème de repérage de concepts est ainsi transformé en un problème d'estimation de l'importance de chaque mot pour un concept donné. Pour un concept donné c associé à n termes $t_1, \dots, t_n$, l'importance du mot w dans le concept c est défini par (Zhou et al., 2006b):

$$I(w) = \max \{ T_j(w) \mid j \leq n \} = \max \left\{ \frac{1/N(w)}{\sum_i 1/N(w_{ji})} \mid j \leq n \right\}$$

avec $T_j(w)$, l'importance du mot w dans le terme t_j,

$N(w)$, le nombre de concepts dont les termes associés contiennent le mot w,

w_{ji}, le i^{ème} token dans le j^{ème} terme associé au concept

L'évaluation de cette méthode sur des textes biomédicaux en utilisant l'UMLS comme dictionnaire montre qu'elle donne des performances supérieures à celles reposant sur la correspondance exacte.

Dans ces approches à base de dictionnaires, les ressources terminologiques utilisées doivent contenir les entrées de tous les concepts cibles, ce qui limite leur portée. Il arrive ainsi qu'elles soient combinées aux méthodes à base de règles pour combler cette limite liée à la couverture du vocabulaire.

```
Phrase: Dementia with Lewy bodies
>>>>> Phrase
dementia with lewy bodies
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
 1000   Dementia with Lewy bodies (Lewy Body Disease) [Disease or Syndrome]
<<<<< Mappings
Phrase: and
>>>>> Phrase
<<<<< Phrase
Phrase: Parkinson's disease dementia
>>>>> Phrase
parkinson disease dementia
<<<<< Phrase
>>>>> Mappings
Meta Mapping (901):
  901   PARKINSON DISEASE (PARKINSON DISEASE (allelic variant)) [Gene or Genome]
  827   Dementia [Mental or Behavioral Dysfunction]
Meta Mapping (901):
  901   PARKINSON DISEASE (PARKINSON DISEASE (allelic variant)) [Gene or Genome]
  827   dementia (Presenile dementia) [Mental or Behavioral Dysfunction]
Meta Mapping (901):
  901   Parkinson Disease [Disease or Syndrome]
  827   Dementia [Mental or Behavioral Dysfunction]
Meta Mapping (901):
  901   Parkinson Disease [Disease or Syndrome]
  827   dementia (Presenile dementia) [Mental or Behavioral Dysfunction]
Meta Mapping (901):
  901   Parkinson's Disease (Parkinson's Disease Pathway) [Functional Concept]
  827   Dementia [Mental or Behavioral Dysfunction]
Meta Mapping (901):
  901   Parkinson's Disease (Parkinson's Disease Pathway) [Functional Concept]
  827   dementia (Presenile dementia) [Mental or Behavioral Dysfunction]
<<<<< Mappings
```

Figure 6 : Exemple de sortie fournie par MetaMap pour la phrase « Dementia with Lewy bodies and Parkinson's disease dementia ». Le chiffre qui précède un concept désigne son score et la chaine de caractères entre crochets son type sémantique.

6.2.2 Les approches à base de règles linguistiques

Des approches alternatives définissant des règles linguistiques (exprimées sous forme d'expressions régulières) permettent d'identifier des termes spécifiques associés à certains types de concepts dans un domaine. Dans (Subramaniam et al., 2003), une méthode combinant la recherche basée sur un dictionnaire et des règles linguistiques est présentée. Par exemple, sachant que beaucoup de noms de protéines terminent en *ase* (*amylase*), les auteurs ont défini une règle permettant de préciser que tout concept ayant une dénomination terminant par le suffixe *ase* est potentiellement une protéine. Des méthodes similaires, basées sur des règles linguistiques, sont proposées dans (Liang et Shih, 2005) et (Wang, 2007) pour la reconnaissance d'entités médicales dénotant des protéines à partir de corpus biomédicaux.

Les principales limites des approches basées sur les règles sont que la mise en œuvre de ces dernières requiert une bonne connaissance du domaine et qu'elles nécessitent un travail manuel généralement lourd et complexe. De plus, les règles sont souvent définies pour un domaine ou une application spécifique et leur application dans d'autres domaines reste problématique.

6.2.3 Les approches statistiques

Les approches statistiques, quant à elles, utilisent généralement des modèles statistiques, tels que les *Conditional Random Fields* (CRF) (Lafferty, 2001) ou les *Hidden Markov Models* (HMMs) (Rabiner, 1989). Ces classifieurs sont entrainés sur des corpus préalablement annotés en sélectionnant des attributs appropriés pour ensuite servir à extraire des concepts à partir de nouveaux documents. Dans la littérature, les approches statistiques ont été largement étudiées. Dans (Xu et al., 2012), les auteurs ont proposé une méthode basée sur les CRF pour extraire des concepts médicaux à partir de textes cliniques. Dans leurs expérimentations, ils ont obtenu de bonnes performances avec une f-mesure de 0.84. He et Kayaalp ont proposé une approche statistique combinant les résultats de MetaMap (y compris les types sémantiques de l'UMLS) avec les CRF pour le repérage d'entités biomédicales (He et Kayaalp, 2008). Settles quant à lui a développé une méthode utilisant les CRF pour la reconnaissance d'entités spécifiques telles que les protéines, les ADN ou les gènes (Settles, 2004). Cette méthode a donné de bons résultats avec une f-mesure de 0,7. Un système de reconnaissance d'entités médicales basé sur les CRF, BANNER[39], a également été proposé dans (Leaman et Gonzalez, 2008). Ce système a été évalué au deuxième challenge BioCreative[40] où il a atteint une f-mesure de 0,85 dans la tâche d'identification des gènes. Le système de reconnaissance d'entités médicales implémenté dans (Campos et al., 2013) donne aussi de bonnes performances avec une f-mesure de 87,2 % sur GENETAG (Tanabe et al., 2005) et une f-mesure de 72,2 % sur JNLPBA (Kim et al., 2004), deux corpus annotés avec des entités de types protéine, ADN et RNA.

Ces méthodes statistiques donnent souvent les meilleures performances, notamment celles utilisant les CRF (Minard et al., 2011; Uzuner et al., 2011) mais nécessitent des donnés annotées manuellement suffisantes pour entrainer les classifieurs. Ces dernières ne sont pas

[39] http://cbioc.eas.asu.edu/banner/
[40] http://biocreative.sourceforge.net/biocreative_2.html

toujours disponibles et leur développement est fastidieux; ce qui est la principale limite de ces méthodes.

6.2.4 Les approches hybrides

Les approches hybrides ont été proposées pour combiner les potentiels des différentes méthodes afin d'améliorer les performances. Ainsi, La combinaison des règles linguistiques avec une approche statistique a été explorée dans (Liang et Shih, 2005) pour l'identification de protéines. Dans (Subramaniam et al., 2003), les auteurs ont montré que les méthodes à base de règles sont complémentaires à celles utilisant des dictionnaires. L'évaluation des différents systèmes d'extraction de concepts participant au challenge *I2B2 2010* (Uzuner et al., 2011) révèle que les performances des méthodes statistiques peuvent être considérablement améliorées par l'exploitation des règles linguistiques et/ou des sources de connaissances externes.

6.3 L'indexation conceptuelle de textes médicaux

Une fois les concepts extraits, ils sont utilisés pour représenter les documents (et les requêtes). Toutefois, les concepts identifiés dans un document ne sont pas tous pertinents pour le représenter; et de plus, ils n'ont pas tous la même importance pour décrire ce dernier. D'où l'intérêt des schémas de pondération et des méthodes de classification qui permettent de déterminer le poids des concepts dans un document (indiquant ainsi leur importance) et de sélectionner les plus représentatifs pour indexer le document.

Nous présentons dans ce qui suit des travaux majeurs dans l'indexation conceptuelle de documents biomédicaux.

6.3.1 Indexation conceptuelle des textes en anglais

Le système MTI (Medical Text Indexer), proposé par la NLM (National Library of Medecine), est l'une des premières initiatives pour l'indexation automatique de documents biomédicaux, spécifiquement les articles scientifiques de la base MEDLINE (Aronson et al., 2004). Ce système repose sur une méthode d'indexation semi-automatique qui utilise les concepts du thésaurus MeSH. Il repose principalement sur MetaMap pour repérer les concepts dans un document. Les résultats de MetaMap sont ensuite combinés à ceux de l'algorithme *PubMed Related Citations* (*PRC*) (Lin et Wilbur, 2007), un algorithme similaire à celui de la recherche des plus proches voisins (*k-NN* (Aha et al., 1991)). De cette combinaison résulte une liste de concepts UMLS, qui est ensuite filtrée pour sélectionner ceux qui correspondent à des concepts MeSH. Enfin, les concepts résultants sont proposés aux experts pour une validation manuelle. Récemment, cette méthode a été étendue avec différentes techniques de filtrage et d'apprentissage automatique pour améliorer ses performances (Mork et al., 2013).

Ruch a quant à lui développé une méthode hybride pour la classification (et donc l'indexation) automatique de textes médicaux (Ruch, 2006). Le système conçu combine deux méthodes. La première est basée sur des règles linguistiques (expressions régulières) pour identifier les concepts dans des textes médicaux. La seconde, quant à elle, utilise les techniques classiques de RI (VSM) en considérant les concepts comme des documents et les

documents comme des requêtes. Ainsi, chaque concept MeSH est indexé comme un document. Ensuite, pour chaque document (requête soumise au SRI), le système retourne une liste ordonnée des concepts MeSH (documents) les plus appropriés. Les résultats de ces deux méthodes sont fusionnés pour fournir une liste finale de concepts avec leurs poids associés. Une évaluation réalisée par les auteurs montre que cette méthode atteint de bonnes performances, comparables aux méthodes basées sur l'apprentissage automatique. Toutefois, elle retourne des résultats bruités avec une tendance à retourner des concepts qui correspondent partiellement aux textes (Trieschnigg et al., 2009).

Les méthodes basées sur l'apprentissage automatique, quant à elles, construisent un modèle à partir d'un ensemble d'entrainement constitué de documents déjà annotés.

Dans (Trieschnigg et al., 2009), les auteurs présentent une étude comparative de six systèmes de classification de documents médicaux en utilisant le thésaurus MeSH. Par des expérimentations, ils ont montré que la méthode basée sur la technique des k-NN (Aha et al., 1991) surpasse les autres, parmi lesquelles figurent le système MTI et l'approche développée dans (Ruch, 2006). Le principe de cette méthode est de considérer les concepts (issus du thésaurus MeSH dans cette étude) attribués aux k documents les plus similaires au document à indexer. Ensuite, ces concepts sont classés par ordre décroissant de leurs scores de pertinence et les top concepts les plus pertinents sont utilisés pour indexer ce document. Dans ce travail (Trieschnigg et al., 2009), la méthode des k-NN est basée sur un modèle de langue (Ponte et Croft, 1998) pour la recherche de documents similaires à un document donné. La pertinence d'un concept pour un document est la somme des scores du SRI associés aux documents annotés par ce concept parmi ses voisins. Les expérimentations de Trieschnigg et ses collègues ont également montré l'utilité des ressources sémantiques pour améliorer les performances de la RI dans le domaine biomédical (Trieschnigg et al., 2009).

Une méthode similaire basée sur la technique des k-NN a été proposée dans (Huang et al., 2011) pour déterminer les voisins d'un document donné. Les concepts assignés à ces derniers sont collectés et classifiés en utilisant un modèle d'apprentissage; dans ce travail, les auteurs ont utilisé un modèle *learning-to-rank* (Liu, 2009). Le principe de ce modèle est de déterminer le score de chaque concept en se basant sur son vecteur d'attributs et de classer les concepts par ordre décroissant de pertinence en utilisant leurs scores. Les concepts ayant les scores les plus élevés sont ensuite sélectionnés. Dans ce travail, les auteurs fixent le nombre de concepts pour indexer un document à 25. Des expérimentations sur deux petits jeux de données standards (200 et 1000 documents) ont montré que cette méthode permet d'obtenir des performances meilleures que le MTI.

Dans (Zhou et al., 2006b), une méthode d'indexation conceptuelle basée sur MaxMatcher (Zhou et al., 2006b) a été proposée. L'utilisation des concepts avec un modèle de langue (plutôt que des mots simples) a permis d'améliorer considérablement les performances de la RI sur la collection de *TREC 2004 Genomics Track* : le MAP (Mean Average Precision) augmente de 29.17% pour la référence (baseline) jusqu'à 36.74% avec cette méthode.

6.3.2 Indexation conceptuelle de textes en français

Pour l'indexation des ressources francophones, Névéol et ses collègues ont proposé le système MAIF (MeSH Automatic Indexing for French) qui combine une méthode de TAL et l'approche des k-NN (Névéol et al., 2006). La méthode de TAL permet d'identifier les concepts MeSH contenus dans les documents en utilisant l'outil INTEX, un environnement de développement linguistique. Quant à la seconde méthode, elle permet d'identifier les documents les plus proches du document à indexer en considérant seulement les titres des documents. Elle utilise pour cela des mesures de similarité telles que la distance de Levenshtein (Levenshtein, 1966).

Dans le même domaine, l'outil F-MTI (French Multi-Terminology Indexer) intègre plusieurs terminologies médicales du portail de santé CISMeF (MeSH, SNOMED , CIM-10 , etc.) pour l'indexation des ressources francophones (Pereira et al., 2009). Il utilise l'algorithme suivant pour le repérage des termes : 1) chaque phrase du texte est normalisée (élimination des accents, utilisation des minuscules) et découpée en mots ; 2) après élimination des mots vides, chaque mot est normalisé en utilisant le stemming ; 3) le *sac de mots* résultant est mis en correspondance avec les entrées des terminologies utilisées sans tenir compte de l'ordre des mots. Cette approche a été particulièrement utilisée pour l'indexation de dossiers médicaux (Pereira et al., 2009).

Une méthode d'indexation conceptuelle est également proposée dans (Dinh et Tamine, 2010) et utilisée pour classifier des documents médicaux à partir du thésaurus MeSH. Les auteurs proposent un mécanisme de désambiguïsation qui exploite la hiérarchie des concepts MeSH pour déterminer le sens adéquat des termes. Les documents sont ensuite indexés par des concepts désambiguïsés. Leurs expérimentations sur une collection de tests standards (TREC9-FT 2000) ont montré une nette amélioration des performances par rapport aux méthodes classiques.

6.4 L'expansion sémantique de requêtes dans le domaine biomédical

Dans le domaine biomédical, les approches de RI sémantique utilisent les concepts issus d'une RTO ou les combinent avec des mots clés pour représenter les documents (et les requêtes). Ensuite, des modèles classiques de RI sont utilisés pour apparier les documents aux requêtes. Cet appariement est déterminant pour les performances d'un SRI. Ainsi, dans le but d'améliorer les performances des méthodes de RI biomédicale, des travaux ont investigué des stratégies d'*expansion de requêtes* (et même de documents). Elles consistent à étendre une requête (ou un document) initiale par des concepts ou termes associés afin d'optimiser la fonction de pertinence du modèle de RI associé. Dans la littérature, différentes techniques ont été proposées pour mettre en œuvre ces stratégies. Certaines approches s'intéressent seulement à l'extension des requêtes (Aronson et Rindflesch, 1997; Díaz-Galiano et al., 2009; Lu et al., 2009; Azcárate et al., 2012) tandis que d'autres ont exploré en plus l'expansion de documents (Diem et al., 2007; Gobeill et al., 2009). Pour cela, des RTO telles que le thésaurus MeSH ou l'UMLS sont couramment utilisées. Dans les différentes campagnes d'évaluation,

les techniques d'expansion de requêtes sont couramment employées et permettent en général d'accroitre les performances.

6.4.1 Expansion de requêtes

Dans (Azcárate et al., 2012), les auteurs ont proposé une méthode d'expansion de requêtes basée sur le thésaurus MeSH. Les termes associés à un descripteur MeSH (cf chapitre 1 section 1.5) sont considérés comme synonymes et utilisés pour étendre les requêtes. Leurs expérimentations montrent que l'expansion de requêtes en utilisant le thésaurus MeSH peut améliorer significativement les performances d'un SRI médical.

Une méthode d'expansion de requêtes similaire est décrite dans (Díaz-Galiano et al., 2009). Cette méthode exploite les liens hiérarchiques du thésaurus MeSH pour étendre les requêtes. L'outil MetaMap (Aronson, 2001) est utilisé pour identifier les concepts de la requête qui sont ensuite complétés par leurs concepts fils. Cette méthode a également donné de bonnes performances sur une collection de tests destinée à la recherche d'images (ImageCLEF 2011) avec une amélioration de 32,11 % de la MAP. Des approches similaires développées dans (Lu et al., 2009; Stokes et al., 2009; Zhou et al., 2007) ont également permis d'améliorer les performances de la RI.

Pour une description plus détaillée et une comparaison des différentes approches d'expansion de requêtes, nous renvoyons le lecteur vers les travaux de (Stokes et al., 2009) et (Bhogal et al., 2007).

6.4.2 Expansion de documents

L'expansion de documents qui consiste à étendre les concepts (ou termes) représentant les documents (par des concepts proches) pour optimiser leur appariement aux requêtes, a aussi été explorée dans le domaine biomédical.

Dans (Gobeill et al., 2009), les auteurs ont présenté une méthode qui combine l'expansion de requêtes et l'expansion de documents en utilisant le thésaurus MeSH. Pour assigner des concepts MeSH aux documents et requêtes, ils ont exploité la méthode de classification de textes proposée dans (Ruch, 2006). Ensuite, les termes dénotant ces concepts sont utilisés pour l'expansion des requêtes et des documents. Enfin, une RI classique est effectuée sur ces documents étendus. L'évaluation de cette méthode sur la collection *ImageCLEF 2008* a montré qu'elle permet d'améliorer considérablement les performances : la MAP croît de 0,14 pour la référence (baseline) jusqu'à 0,18 (soit une amélioration de 29 %) avec cette méthode.

(Diem et al., 2007) ont également exploré l'expansion de requêtes et de documents en exploitant les relations sémantiques de l'UMLS. Les auteurs ont utilisé les relations taxonomiques directes pour étendre les requêtes par des concepts plus spécifiques (enfants directs) et les documents par des concepts plus génériques (parents directs). Ils ont rapporté dans leurs expérimentions une nette amélioration des résultats : une augmentation de la MAP de 66 % comparativement à une recherche par mots clés et de 34 % par rapport à une indexation conceptuelle sans expansion.

Bien que l'expansion de requêtes (et documents) permette d'améliorer les performances en RI, des travaux ont prouvé qu'elle pouvait aussi les dégrader (Hersh et al., 2000). Les résultats dépendent d'un ensemble de variables (Hersh et al., 2007; Dinh, 2012) : les ressources sémantiques utilisées, la méthode d'extraction de concepts, le modèle de RI, la stratégie utilisée pour l'expansion, etc.

7 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art de différents travaux sur la RI en nous focalisant sur les approches sémantiques.

Dans un premier temps, nous avons exploré les différents champs de la RI, d'une manière générale. Ainsi, après avoir décrit le principe de fonctionnement d'un SRI, nous avons présenté les modèles de RI les plus courants : le modèle booléen, le modèle vectoriel et le modèle probabiliste. Une variété de mesures utilisées et les campagnes pour évaluer les performances des SRI ont aussi été exposées.

Dans un deuxième temps, nous nous sommes intéressés aux travaux sur la RI sémantique. Ces derniers peuvent être divisés en deux catégories : les approches statistiques basées sur la sémantique intrinsèque du contenu des documents et les approches utilisant des ressources sémantiques externes. La deuxième approche, bien que prometteuse, nécessite des méthodes appropriées pour le traitement d'un certain nombre de tâches. En effet, dans cette approche, les documents (et les requêtes) sont représentés par des concepts (indexation conceptuelle) plutôt que par des mots clés. Ce type d'indexation nécessite, au préalable, l'extraction des concepts dans des documents et leur éventuelle désambiguïsation. L'expansion de requête a été aussi investiguée afin d'améliorer les performances de la RI. Bien que la plupart de ces travaux utilisent des ressources générales comme WordNet, d'autres ont exploité des ressources sémantiques spécifiques telles que l'UMLS. On note ainsi des travaux majeurs dans le domaine biomédical où beaucoup de méthodes et outils ont été développés. Ceci s'explique non seulement par l'abondance des RTO dans le domaine mais aussi par l'explosion des données.

Bien que d'importantes avancées aient été notées dans la RI sémantique biomédicale, il y a encore des défis (identification des concepts, désambiguïsation éventuelle, sélection des concepts pertinents, expansion des requêtes, couverture limitée des ressources) que nous abordons dans les chapitres 4 et 5.

Le chapitre suivant présente la méthodologie que nous proposons pour construire une ontologie de domaine permettant de mettre en œuvre dans un deuxième temps un modèle de RI sémantique.

Chapitre 3: Réutilisation de ressources de connaissances existantes pour la construction d'ontologies

1 Introduction

L'engouement suscité par les multiples fonctionnalités offertes par les ontologies a résulté en d'importants travaux qui ont abouti au développement de nombreuses ressources de connaissances, notamment dans le domaine biomédical. Comme souligné dans le Chapitre 1, ces ressources sont de différents types et concernent aussi bien des connaissances générales (e.g. UMLS) que des connaissances spécifiques à un sous-domaine de la médecine (e.g., le FMA). Toutefois, bien que l'abondance de RTO soit une réalité dans ce domaine, il est malgré tout nécessaire de tenir compte de l'évolutivité de la connaissance dans des domaines particuliers comme celui de la maladie d'Alzheimer où les connaissances sont régulièrement enrichies. La modélisation des connaissances sur cette maladie et ses syndromes associés reste aujourd'hui un enjeu important pour une meilleure compréhension de la maladie afin d'aider à une prise de décision éclairée en termes de prévention, de prise en charge, etc.

La conception d'ontologie est une tâche fastidieuse et coûteuse, qui nécessite généralement beaucoup de ressources (temps, expertise). Nous l'avons largement abordée dans le chapitre 1 de ce manuscrit. Pour simplifier le processus ou tout au moins alléger certaines tâches, l'acquisition (semi-)automatique d'ontologies à partir de textes a été proposée comme une alternative. Depuis, plusieurs travaux qui suivent cette voie ont vu le jour (par exemple (Cimiano et Völker, 2005) et (Aussenac-Gilles et al., 2008)). L'acquisition d'ontologies à partir de textes exploite des méthodes et outils de TAL. Cette manière de construire des ontologies part du postulat que les textes sont des sources de connaissances importantes, en particulier dans le domaine biomédical. En effet, ce domaine est particulièrement riche en documents textuels (guides de bonnes pratiques, comptes rendus médicaux, articles scientifiques, etc.). Toutefois, l'acquisition d'ontologie à partir de textes reste généralement confrontée aux problèmes déjà énoncés au chapitre 1 (conceptualisation manuelle fastidieuse, beaucoup de bruit avec la conceptualisation automatique, incomplétude des textes). Pour y répondre, une nouvelle approche s'est intéressée à la réutilisation de ressources existantes. Dans le domaine biomédical, nous pouvons mentionner les travaux de Hahn et Schulz sur la formalisation d'une partie de l'UMLS (Hahn et Schulz, 2004), de Jiménez-Ruiz et ses collègues sur la réutilisation du thésaurus NCI (National Cancer Institute) et de l'ontologie GALEN pour construire une ontologie de l'arthrite chronique juvénile (Jiménez-Ruiz et al., 2008) ou encore de Charlet et ses collègues sur l'exploitation de RTO existantes pour la mise en œuvre de l'ontologie *ONTOLURGENCES* (Charlet et al., 2012). Cette approche propose d'exploiter partiellement ou entièrement les contenus de ces ressources existantes, qui peuvent être implicites (niveau de spécification faible) et/ou informelles (codées dans des langages non formels), afin de construire des ontologies.

Bien que des avancées majeures aient été réalisées (cf section 7.3, chapitre 1) dans ce cadre (développement d'outils supportant cette tâche, des expérimentations dans différents domaines, etc.), des challenges tels que l'explicitation, la formalisation et la couverture limitée des ressources de connaissances se posent toujours.

C'est ainsi que le travail que nous décrivons dans ce chapitre vise à proposer une méthodologie permettant de résoudre ces différentes questions. Nous présentons une approche de construction d'une ontologie de domaine, dans le domaine médical, à partir de ressources existantes. L'approche est décrite à travers une application concrète : la mise en œuvre d'une ontologie bilingue de la maladie d'Alzheimer et les syndromes apparentés. Bien que des ontologies spécifiques du domaine aient été développées (Malhotra et al., 2012; Jensen et al., 2013), elles ne couvrent pas suffisamment les différents aspects de la maladie traités dans la base BiblioDem. En plus, elles sont monolingues (anglais) et sont donc limitées pour l'objectif de ce travail. Nous proposons ainsi d'illustrer notre méthodologie pour la construction d'une ontologie bilingue qui couvre spécifiquement et suffisamment cette maladie. Cette ontologie sera utilisée pour supporter un portail sémantique de RI.

Dans un premier temps, et en guise de préliminaire, nous définissons les différentes notions utilisées tout au long de ce chapitre (section 2). Nous présentons ensuite l'architecture générale de la méthodologie proposée (section 3) avant d'illustrer ses différentes étapes sur le cas spécifique de la maladie d'Alzheimer (section 5). Nous exposons les différentes ressources que nous avons utilisées (section 4). Dans la section 6, l'ontologie de la maladie d'Alzheimer est détaillée et des aspects méthodologiques sont discutés. Nous terminons par une conclusion et des perspectives à envisager pour cette partie de notre travail.

2 Définitions des notions de base

Définition 1 : En adaptant la définition donnée par Maedche et Staab (Maedche et Staab, 2001), nous définissons formellement une ***ontologie*** comme un 8-uple $O = \{ C, H, R, H_R, T, F, G, A\}$ où :

- C est un ensemble de concepts ;
- H $\subseteq$ C x C est une taxinomie de concepts. $h = (c_1, c_2) \in$ H signifie que c_1 est subsumé par c_2 ;
- R $\subseteq$ C x C x L est un ensemble de relations associant deux concepts où L est un ensemble de labels de relations ;
- $H_R \subseteq$ R x R est une taxonomie de types de relations transversales ;
- T est l'ensemble des termes dénotant les concepts ;
- F : T $\rightarrow$ C est une fonction qui associe les termes aux concepts qu'ils dénotent ;

- $G : C \rightarrow T$ est une fonction qui associe les concepts à leurs termes. Les fonctions F et G sont inverses ;
- A est un ensemble d'axiomes permettent de décrire des contraintes dans l'ontologie.

Définition 2 : H^+ est la ***fermeture transitive*** de H si et seulement si H^+ est le plus petit ensemble de relations tel que H^+ est transitive et $H \subseteq H^+$.

Définition 3 : Un ***syntagme nominal*** est une séquence de tokens, où l'élément principal a pour catégorie grammaticale « nom » (autrement dit, un groupe nominal).

Définition 4 : Un ***terme*** est une unité terminologique qui désigne une notion précise dans un domaine donné et qui est constituée d'un mot - on parle alors de terme simple - ou de plusieurs mots - on parle dans ce cas de terme complexe ou multi-mots.

Définition 5 : Un ***candidat terme*** est un mot ou une suite de mots susceptibles d'être un terme, c'est-à-dire une entrée d'une RTO (Bourigault et Aussenac-Gilles, 2003).

Définition 6 : Un ***token*** est une chaîne de caractères alphanumériques qu'un programme informatique peut reconnaitre. Un mot est un token ayant une signification. Par exemple, « vitamine », « B12 », « traitement » sont des tokens parmi lesquels « vitamine » et « traitement » sont aussi des mots.

Définition 7 : Un ***corpus parallèle***, appelé aussi bitexte, est un ensemble de textes dans deux langues différentes dont les uns sont des traductions des autres.

Nous allons à présent décrire l'approche TOReuse2Onto de construction d'une ontologie associée à une terminologie bilingue, en se basant sur l'extraction de termes au sein de corpus et la réutilisation de ressources existantes.

3 L'approche TOReuse2Onto

La méthodologie que nous proposons intègre une variété de techniques allant des méthodes de TAL pour l'extraction d'informations à partir de textes à la réutilisation de RTO existantes pour structurer ces informations. Comme illustrée dans la figure 7, elle est inspirée de la méthodologie Terminae (Aussenac-Gilles et al., 2008). Nous y avons intégré la réutilisation de RTO dans la phase de conceptualisation, et surtout, une étape d'alignement de termes pour tenir compte de l'aspect multilingue de notre approche. TOReuse2Onto comprend cinq étapes : 1) la constitution de corpus, 2) l'extraction des candidats termes dans les corpus, 3) la construction du noyau ontologique, 4) l'enrichissement de l'ontologie par l'intégration de nouveaux concepts et l'alignement de termes dans des langues différentes et enfin 5) la validation de l'ontologie par des experts du domaine.

Après une brève description des différentes étapes, nous présentons en détail la manière dont nous les avons mises en œuvre dans le cadre de la construction de l'ontologie sur la maladie d'Alzheimer.

La constitution de corpus

Cette étape consiste à construire un corpus sur lequel des méthodes de TAL sont appliquées pour extraire les termes et les relations lexicales et syntaxiques entre ces termes. Ce corpus doit être représentatif afin de couvrir au maximum les connaissances du domaine d'application cible. Il doit être également d'une qualité suffisante pour permettre l'extraction d'entités (termes, relations) pertinentes du domaine. Plusieurs types de textes peuvent être utilisés pour construire le corpus, comme par exemple des documents techniques ou encore des résumés d'articles scientifiques publiés dans le domaine.

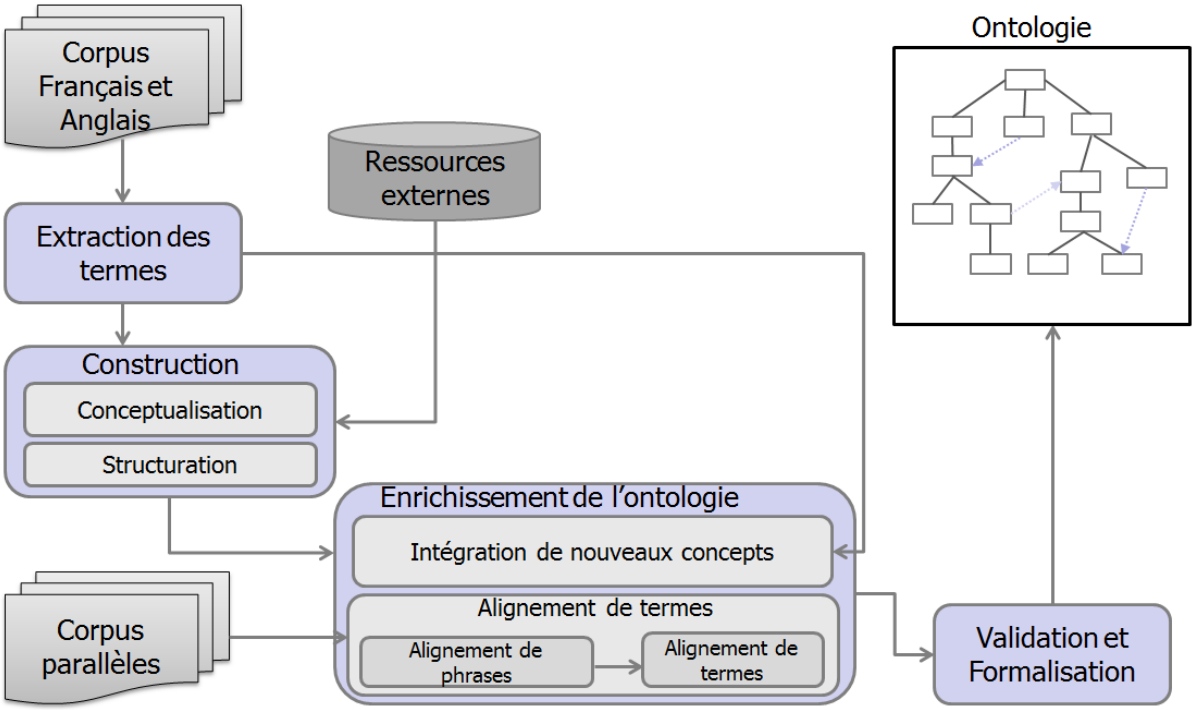


Figure 7 : Architecture générale de notre méthodologie de construction d'ontologie : TOReuse2Onto

L'extraction des candidats termes

Cette étape assure l'analyse des textes pour extraire les candidats termes du domaine et les relations lexicales ou syntaxiques qui les lient. Dans la littérature, une large gamme de méthodes et outils de TAL ont été développés pour supporter l'extraction terminologique à partir de textes. On peut les classer en trois catégories : les approches statistiques, les approches linguistiques et les approches mixtes. Les approches linguistiques s'intéressent à des candidats termes ayant une certaine structure grammaticale, comme les noms (*maladie*, *traitement*), les syntagmes nominaux (*maladie infectieuse*, *corps de Lewy*), les syntagmes adjectivaux (*très sévère*), etc. Les approches statistiques exploitent les mesures statistiques,

telles que la fréquence, le poids relatif (mesuré par le TF.IDF) ou l'entropie, pour extraire les candidats termes représentatifs d'un corpus. Généralement, ce sont les approches hybrides, combinant des techniques de ces deux catégories, qui sont utilisées.

La construction du noyau ontologique

Cette étape consiste à conceptualiser les termes extraits dans la phase précédente et à structurer les concepts résultants grâce à des relations sémantiques (taxonomiques et non taxonomiques). Pour ce faire, des ressources sémantiques existantes sont exploitées afin d'identifier les concepts correspondant aux termes sélectionnés, puis extraire les informations sur ces derniers et les structurer. Cette étape nécessite une série de traitements dont l'alignement des termes aux entrées des ressources de connaissances ciblées et la suppression d'éventuelles redondances et incohérences.

L'enrichissement de l'ontologie

Cette étape permet d'enrichir l'ontologie obtenue lors de la phase précédente en y intégrant de nouvelles entités. Ces dernières peuvent être des termes, des concepts et/ou des relations sémantiques. Dans le cadre de la construction d'une ontologie multilingue, comme c'est notre cas, elle peut traiter l'alignement de termes de langues différentes. Dans le cas spécifique de notre application, elle comprend deux phases : l'**alignement de termes** et l'**intégration de nouveaux** concepts en exploitant les dépendances syntaxiques entre termes. Pour la première, nous proposons une méthode d'alignement terminologique basée sur des corpus parallèles. Pour la deuxième, les relations syntaxiques fournies par un analyseur sont exploitées. Les techniques proposées dans la littérature pour établir des relations taxonomiques entre les concepts, qui sont présentées dans le chapitre 1, peuvent être utilisées alternativement ou en complément pour cette tâche.

La validation et la formalisation de l'ontologie

C'est la dernière étape qui fait appel aux experts du domaine pour vérifier et valider l'ontologie résultante. Elle permet de résoudre des incohérences, de raffiner l'ontologie et d'élaguer ses éléments non pertinents mais aussi d'évaluer sa couverture du domaine. Après validation, l'ontologie est décrite dans un langage formel, tel qu'OWL.

Dans la suite de ce chapitre, nous allons revenir en détail sur ces différentes étapes en les illustrant sur le cas d'usage de la maladie d'Alzheimer. Nous présentons tout d'abord les différentes ressources et outils spécifiques utilisés pour ce cas d'usage.

4 Ressources utilisées

Dans l'approche de construction d'ontologies que nous avons mise en œuvre, différents outils et ressources peuvent être utilisés. Ils ne font pas partie intégrante de la méthodologie car ils peuvent être remplacés par des outils similaires. Nous présentons dans un premier temps la ressource documentaire que nous avons utilisée pour constituer les corpus de textes : la base de données du bulletin BiblioDémences. Ensuite, nous décrivons la base de connaissances

médicales que nous avons exploitée pour organiser les termes entre eux. Enfin, nous décrivons les outils de TAL que nous avons utilisés : Syntex (Bourigault et Fabre, 2000) pour l'extraction de candidats termes et Moses (Koehn et al., 2007) pour l'alignement de termes anglais-français pour l'aspect bilingue de la ressource à construire.

4.1 BiblioDémences

Le bulletin bibliographique BiblioDémences[41], lancé depuis janvier 2004 par l'équipe *Epidémiologie et Neuropsychologie du Vieillissement Cérébral*[42] du centre INSERM U897, fournit des analyses critiques d'articles scientifiques portant sur la maladie d'Alzheimer et les syndromes apparentés. À partir d'une veille bibliographique de la littérature mondiale (dans de vastes bases documentaires, telles que MEDLINE), une trentaine d'articles sont sélectionnés manuellement chaque mois et proposés à des spécialistes du domaine pour qu'ils en fassent une analyse critique. Cette dernière comprend les éléments suivants : i) une traduction du titre en français ; ii) une synthèse du contenu de l'article (mais qui n'est pas une traduction du résumé) ; iii) des commentaires sur l'article. Le comité éditorial de BiblioDémences consulte ensuite ces analyses et en sélectionne une dizaine pour parution dans le bulletin bibliographique qui parait régulièrement. Tous les articles analysés (sélectionnés ou non pour publication dans BiblioDémences) sont intégrés dans la base de données BiblioDem avec la lecture critique associée. La figure 8 montre un exemple d'analyse critique d'un article.

La maladie d'Alzheimer

Synthèse La MA est la principale cause de démence. L'avancée des recherches a permis une compréhension détaillée au niveau moléculaire de la maladie (plaques séniles composées d'amyloïde beta (A-beta), enchevêtrements neurofibrillaires dus à l'hyperphosphorylation de Tau). Pourtant, au fur et à mesure de l'augmentation des connaissances, on se rend compte de la complexité de cette pathologie. La forme familiale de la MA est une maladie autosomale dominante rare, qui débute précocement au cours de la vie. Elle est causée par des mutations des gènes de la protéine précurseur de l'amyloïde (APP) et des présenilines (PSEN 1 et 2) qui sont liés au métabolisme de l'A-beta. La forme sporadique est au contraire plus fréquente et touche environ 15 millions de personnes à travers le monde. Les causes de la maladie sous sa forme sporadique ne sont pas toutes connues, du fait de l'hétérogénéité de la maladie, mais, on sait que l'âge et les complexes interactions des facteurs de risque à la fois génétiques et environnementaux sont en partie responsables de son apparition. Cette revue expose les aspects clés de la maladie incluant les données épidémiologiques, génétiques, moléculaires, le diagnostic, les traitements ainsi que les avancées récentes et les controverses. En perspectives, les auteurs soulignent qu'il reste à vérifier l'hypothèse de la cascade amyloïde (toxicité de l'A-beta, cause ou conséquence de la neurodégénérescence ?) et à déterminer le rôle du dysfonctionnement axonal. En réponse aux traitements, on devrait observer une diminution des plaques séniles, de la charge d'A-beta, et une amélioration des capacités cognitives. Les modèles murins surestiment souvent les bénéfices de nouvelles stratégies de traitement et l'application des résultats à l'homme doit se faire avec prudence. Le futur repose sur les avancées en recherche fondamentale, dans le développement des techniques et dans les progrès de la recherche clinique dans le but de développer à terme une vraie stratégie de prévention de la MA.

Commentaires Excellente synthèse sur la MA, argumentée et riche de schémas explicatifs. A lire !

Analysé par Catherine Féart, CMRR Aquitaine

Figure 8 : Exemple d'analyse critique d'un article de la base BiblioDem

[41] http://sites.isped.u-bordeaux2.fr/bibliodem/bulletins.aspx
[42] http://www.isped.u-bordeaux2.fr/CDD/FR_HTM_equipe.aspx?CLE_EQU=3

BiblioDem est une base cumulative qui contenait, en juillet 2014, environ 1 900 articles scientifiques sur la maladie d'Alzheimer et les syndromes apparentés.

4.2 L'UMLS

L'UMLS®[43] (Unified Medical Language System) (Lindberg et al., 1993; Bodenreider, 2004) est une des bases de connaissances les plus importantes dans le domaine biomédical. Elle est développée par la *National Library of Medicine*[44] (NLM) et a pour objectif l'intégration d'une variété de RTO (les « vocabulaires sources ») au sein d'un système unifié. L'UMLS est composé de trois éléments principaux : le réseau sémantique, le Metathesaurus et le lexique spécialiste.

4.2.1 Le réseau sémantique

Le réseau sémantique de l'UMLS comprend une hiérarchie de 133 types sémantiques organisés suivant 49[45] types de relations. Ces types sémantiques, qui permettent de catégoriser les concepts du Metathesaurus, sont structurés selon des relations hiérarchiques et des relations transversales. Les hiérarchies sont constituées par deux types de relations : la relation de subsomption (définie par *is_a*) et la relation méronymique (définie par *part_of*). Par exemple, *Organism*, *Anatomical Structure* et *Substance* sont des sous classes de *Physical Object* (au sens de la relation de subsomption). Les relations transversales, quant à elles, concernent une variété de liens sémantiques : *affects*, *associated with*, *complicates*, *location_of*, *co-occurs with*, etc. Par exemple, la relation *affects* est définie entre les types sémantiques *Anatomical Abnormality* et *Organism*. Les types sémantiques ont été regroupés en quinze groupes sémantiques qui correspondent à des catégories générales du domaine biomédical (ex : ***Anatomy***, ***Disorders***, etc.) (Bodenreider et al., 2003).

4.2.2 Le Metathesaurus

Le Metathesaurus® est un large graphe intégrant actuellement 161 vocabulaires sources couvrant différentes spécialités du domaine biomédical (Figure 9). Il est constitué de plus de deux millions de nœuds (concepts) structurés par plus de 50 millions de relations hiérarchiques et transversales. Les concepts regroupent des termes synonymes provenant des divers vocabulaires sources, représentant ainsi un sens unifié pour l'ensemble des termes. Chaque concept a un identifiant unique, appelé CUI (Concept Unique Identifier), et est associé à plusieurs termes le dénotant. Un terme possède lui aussi un identifiant unique (LUI - Lexical Unique Identifier) et est associé à ses différents variants lexicaux. Ces variants sont générés en utilisant le *Lexical Variant Generation*[46]. Un concept possède également une ou plusieurs définitions. Les termes, tout comme les définitions, sont en anglais et il existe parfois des équivalents dans d'autres langues, dont le français. Enfin, chaque concept du Metathesaurus est catégorisé par un ou plusieurs types sémantiques du réseau sémantique (par exemple, le concept *Alzheimer's disease* a pour type sémantique *Disease or Syndrome*).

[43] http://www.nlm.nih.gov/research/umls/
[44] http://www.nlm.nih.gov/
[45] Les chiffres fournis dans ce document concernent la version 2012AA de l'UMLS
[46] http://www.nlm.nih.gov/research/umls/new_users/online_learning/LEX_004.htm

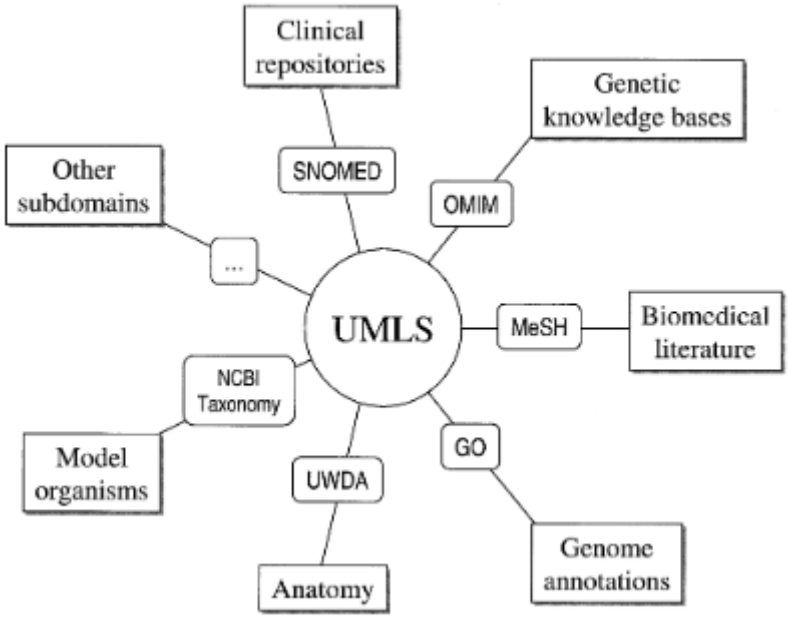


Figure 9 : Exemple de ressources intégrées dans l'UMLS avec leurs domaines d'application (Bodenreider, 2004)

4.2.3 Le lexique spécialiste

Le lexique spécialiste[47] recense des informations lexicales, orthographiques et syntaxiques sur les termes qui sont associés aux concepts. Il intègre une bonne partie des termes biomédicaux anglais avec leurs variations morphosyntaxiques. Pour chaque terme, le lexique contient sa forme normalisée (appelée « lemme ») et ses différents variants. Par exemple, *neurodegenerative disorder* est un variant orthographique de *Neurodegenerative Disorders*. Le lexique spécialiste est une ressource libre d'accès et open-source.

En plus de ces trois composants, l'UMLS est associé à un ensemble d'outils facilitant son exploitation. En particulier, les services *UTS*[48] (UMLS Terminology Services) sont des outils permettant de déterminer quel(s) concept(s) correspond(ent) à un terme fourni en entrée suivant diverses options de recherche : en fonction des vocabulaires sources ciblés, des types sémantiques spécifiés, etc.

4.3 Les outils utilisés

4.3.1 Syntex

Syntex est un analyseur syntaxique qui permet d'extraire des candidats termes (noms, syntagmes nominaux) et des relations syntaxiques les structurant à partir de corpus de textes (Bourigault et Fabre, 2000). A partir d'un texte étiqueté par un outil comme TreeTagger, Syntex fournit un réseau de dépendance constitué de mots et de syntagmes (nominaux, verbaux, adjectivaux) avec leurs contextes d'apparition. Dans ce réseau terminologique,

[47] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html
[48] https://uts.nlm.nih.gov/home.html

chaque syntagme est relié à sa tête et à son expansion syntaxique. Par exemple, le syntagme nominal *déclin cognitif léger* a pour tête *déclin cognitif* et pour expansion syntaxique *léger*. Syntex, comme son prédécesseur Lexter (Bourigault, 1994), intègre une technique d'apprentissage endogène qui lui permet de *lever les ambiguïtés de rattachement syntaxique*. Le principe de cet apprentissage est de repérer dans le corpus des occurrences des syntagmes concernés non ambigües pour résoudre les problèmes d'ambiguïté syntaxique.

La chaîne de traitement commence par une identification des dépendances entre les mots d'une phrase (par exemple, sujet d'un verbe, complément d'un nom). Ensuite, à partir de ces dépendances, l'outil génère les syntagmes maximaux qui seront, par la suite, analysés et décomposés en tête et expansion. Ainsi, pour la phrase « *Cette étude porte sur les facteurs de risques de démence*. », le syntagme nominal maximal *facteurs de risques de démence* est généré avant d'être décomposé en *facteurs de risques* et *démence*. Chaque syntagme est associé à un ensemble de mesures que sont sa fréquence (son nombre d'occurrences dans le corpus), et sa productivité (son utilisation dans des termes plus complexes) qui permet de déterminer les syntagmes partageant les mêmes têtes ou les mêmes expansions.

Des outils similaires d'extraction terminologique tels que YaTeA (Aubin et Hamon, 2006) et ACABIT (Daille, 1994) permettent aussi de traiter des corpus volumineux.

4.3.2 Moses

Moses[49] (Koehn et al., 2007) est un système de traduction automatique statistique (*Statistical Machine Translation*) à base de segments (séquences continues de mots ou n-grammes), open source et largement utilisé. Il consiste en un ensemble de composants qui permettent de prétraiter des données, d'entrainer des modèles de langues et des modèles statistiques de traduction pour toute paire de langues. Ainsi, à partir d'un corpus parallèle prétraité préalablement (segmentation en mots normalisation des mots, des modèles de traduction sont générés en utilisant un composant dédié à cette tâche. Pour cela, Moses repose sur GIZA++ (Och et Ney, 2003), un des outils d'alignement de mots les plus connus, pour établir des alignements entre mots et utilise les cooccurrences des mots et des segments de mots afin de produire des correspondances entre les segments. Il permet également de prendre en compte des informations linguistiques, telles que la forme normalisée (lemme) et la catégorie grammaticale des mots dans des modèles pour améliorer la qualité de la traduction. Moses permet donc de produire, de façon automatisée, à partir de corpus parallèles, des tables de traduction dans lesquelles les paires de segments sont associées à leurs probabilités de traduction. A la différence de la plupart des systèmes d'alignement, Moses permet d'aligner des segments et, par conséquent, des termes multi-mots. Pour chaque couple de segments alignés, il calcule les grandeurs suivantes :

- $p(e|f)$, la probabilité de traduction du segment e sachant le segment f ;
- $p(f|e)$, la probabilité de traduction du segment f sachant le segment e ;

[49] http://www.statmt.org/moses/

- des poids lexicaux pour indiquer le degré de correspondance entre les mots les constituant.

Le tableau 1 synthétise les ressources et outils présentés dans cette section et précise leurs différents rôles dans le processus de construction de l'ontologie détaillé dans la section suivante.

Tableau 1 : Les différentes ressources et outils utilisés et leurs rôles

Ressources/outils	Etape d'utilisation	Rôles
BiblioDem	Constitution des corpus	Une base de données bibliographique contenant des articles spécifiques à la maladie d'Alzheimer, utilisée pour constituer les corpus du domaine (4.1).
L'UMLS	Structuration des termes en concepts dans l'ontologie	Large base de connaissances biomédicales utilisée pour regrouper et lier les termes biomédicaux extraits dans les corpus (4.2).
Syntex	Extraction de termes candidats	Analyseur syntaxique utilisé pour l'extraction de termes et de relations syntaxiques à partir des corpus textuels (4.2.1).
Moses	Enrichissement de l'ontologie	Système de traduction automatique utilisé pour l'alignement des termes à partir de corpus parallèles (4.2.2).

5 Illustration pour la construction d'une ontologie de la maladie d'Alzheimer

5.1 La constitution des corpus

Afin de gérer l'aspect bilingue de la ressource à construire, nous avons constitué deux corpus textuels à partir de la base bibliographique BiblioDémences : un corpus anglais composé des titres et des résumés des articles et un corpus français regroupant les titres traduits, les synthèses et les commentaires de ces mêmes articles. Le tableau 2 présente le nombre de phrases et de mots constituant ces corpus.

Tableau 2 : Statistiques des corpus utilisés

	Corpus français	Corpus anglais
Nombre de phrases	31 595	18 406
Nombre de mots	783 853	375 431

5.2 L'extraction des candidats termes

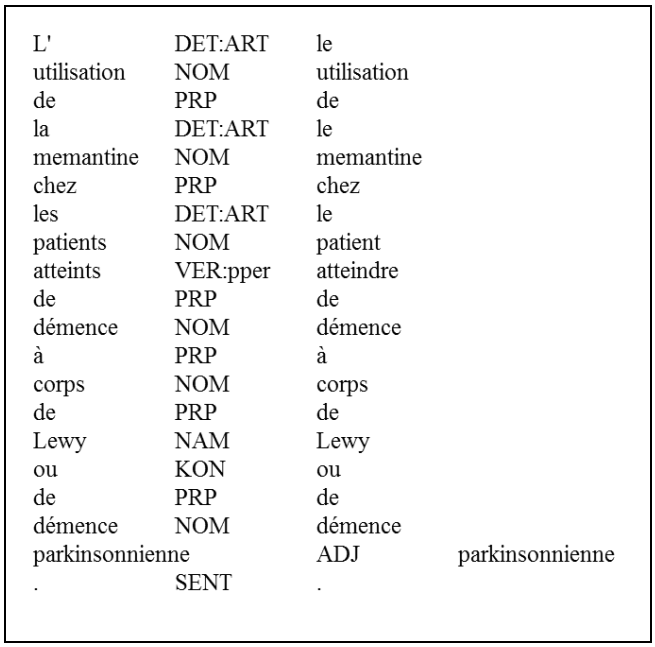
Avant de pouvoir utiliser l'outil Syntex pour extraire les candidats termes à partir de ces corpus, le texte doit être préalablement traité par un étiqueteur grammatical (appelé aussi *étiqueteur morphosyntaxique* ou *part-of-speech (POS) tagger* en anglais).

5.2.1 Prétraitement du texte

L'étiquetage grammatical consiste à associer aux mots du texte des informations grammaticales, telles que leur nature (nom, adjectif, verbe, article, etc.) et éventuellement leur forme canonique. Pour étiqueter nos corpus, nous avons utilisé TreeTagger (Schmid, 1994) qui est l'un des étiqueteurs les plus couramment utilisés pour réaliser cette tâche. Il repose sur une méthode d'étiquetage qui utilise les arbres de décisions pour déterminer ces informations. TreeTagger[50] supporte plusieurs langues (français, anglais, allemand, etc.) et est adaptable à d'autres langues si un lexique et un corpus d'entrainement annoté manuellement sont disponibles pour ces dernières. A titre d'illustration, la figure 10 présente une sortie de TreeTagger.

La première colonne désigne le mot original, la deuxième sa nature et la troisième sa forme canonique (appelée aussi lemme). Il convient de noter qu'un outil similaire, le *Stanford Tagger*[51], a été développé par une équipe de l'université de Stanford pour supporter l'étiquetage morphosyntaxique (Toutanova et al., 2003). En plus de ses bonnes performances, il supporte lui aussi plusieurs langues. Une variété d'étiqueteurs est proposée dans la littérature, souvent associés à d'autres outils de TAL : *OpenNLP tagger*, *Genia Tagger*, *Lingpipe tagger* etc.

[50] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[51] http://nlp.stanford.edu/software/tagger.shtml



L'	DET:ART	le	
utilisation	NOM	utilisation	
de	PRP	de	
la	DET:ART	le	
memantine	NOM	memantine	
chez	PRP	chez	
les	DET:ART	le	
patients	NOM	patient	
atteints	VER:pper	atteindre	
de	PRP	de	
démence	NOM	démence	
à	PRP	à	
corps	NOM	corps	
de	PRP	de	
Lewy	NAM	Lewy	
ou	KON	ou	
de	PRP	de	
démence	NOM	démence	
parkinsonnienne		ADJ	parkinsonnienne
.	SENT	.	

Figure 10 : Illustration de la sortie de TreeTagger pour la phrase « *L'utilisation de la mémantine chez les patients atteints de démence à corps de Lewy ou de démence parkinsonnienne.* »

5.2.2 Extraction des candidats termes

Pour extraire les candidats termes à partir des corpus de départ, nous avons donc choisi l'outil Syntex (Bourigault et Fabre, 2000) car c'est un analyseur syntaxique robuste qui, en plus de supporter le traitement des corpus français et anglais, décrit des relations de dépendance syntaxique entre les différents candidats termes (cf section 4.3.1). Parmi les résultats fournis par Syntex, nous avons sélectionné uniquement les noms et syntagmes nominaux du réseau, considérés comme les candidats termes. L'idée est que généralement les termes sont représentés par des noms communs et des groupes nominaux. Par conséquent, cibler les mots et groupes de mots ayant ces structures syntaxiques permet d'extraire les termes essentiels du corpus mais aussi de limiter le bruit que peuvent engendrer les non-termes.

5.2.3 Filtrage des résultats

La liste de candidats termes nécessite d'être filtrée pour obtenir un résultat plus propre. Ainsi, nous avons sélectionné les candidats termes les plus représentatifs, i.e., ayant une fréquence d'apparition dans le corpus supérieure à un seuil déterminé en accord avec des experts de la maladie d'Alzheimer. Après consultation de la liste de termes, ces experts ont jugé suffisant de s'intéresser aux termes ayant au moins sept occurrences dans le corpus. Par ailleurs, les termes constitués exclusivement de chiffres sont élagués. Ceux ayant moins de quatre caractères et ceux contenant des chiffres ou des caractères non alphanumériques sont isolés pour être présentés aux experts qui décideront s'il faut, ou non, les garder.

A côté de la fréquence, d'autres mesures sont utilisées pour estimer la pertinence des termes pour un corpus de textes. Pour une description plus détaillée de ces mesures, nous renvoyons le lecteur vers les travaux de (Velardi et al., 2001).

Une fois les candidats termes extraits et filtrés, il est ensuite nécessaire de les organiser en concepts et de relier ces derniers entre eux.

5.3 La construction du noyau ontologique

Cette étape consiste à regrouper les termes synonymes en concepts (5.3.1) et à les organiser via des relations sémantiques (5.3.2).

5.3.1 Conceptualisation

La conceptualisation manuelle demeurant une tâche coûteuse et fastidieuse, l'UMLS a été exploité pour automatiser cette tâche. Nous avons ainsi utilisé le service Web de l'UMLS[52] qui permet de retrouver le(s) concept(s) du Metathesaurus correspondant à un terme donné. En effet, dans le Metathesaurus, les termes dénotant une même « notion » sont regroupés dans un concept et une recherche avec un de ses termes permet de retrouver ce dernier. Dans un premier temps, une recherche exacte a été effectuée sur la liste de tous les candidats termes sélectionnés dans l'étape précédente et pour ceux qui n'ont pas été retrouvés tels quels lors de cette première étape, une recherche normalisée a été réalisée. La méthode de recherche normalisée prend en compte les variations lexicales des termes (flexion, dérivation) en se basant sur le programme *Lexical Variant Generation*. Le candidat terme est d'abord normalisé en un ensemble de variants qui sont ensuite comparés aux termes des concepts du Metathesaurus. Ainsi, la recherche normalisée avec les termes *Parkinson's disease dementia* et *Older adults* permet de retrouver respectivement les concepts *Dementia in Parkinson's disease* (CUI C0349081) et *Older Adulthood* (C1999167). Ce traitement avec le LVG est réalisé spécifiquement pour les termes anglais. Concernant les termes français, TreeTagger est utilisé pour les lemmatiser (i.e. les normaliser).

Pour chaque concept retrouvé dans le Metathesaurus de l'UMLS, ses éventuels synonymes et définitions (anglais et français) ont été récupérés. Pour chaque terme synonyme, nous considérons seulement sa forme préférée (cette information est fournie dans l'UMLS) pour éviter d'intégrer trop de termes avec les variants. En plus de la version française de la SNOMED 3.5, le portail terminologique de santé de CISMeF (Tayeb et al., 2011) qui intègre plusieurs ressources terminologiques en français a été aussi exploité pour enrichir l'ontologie avec des synonymes et définitions en français; les termes (normalisés) sont automatiquement alignés aux entrées des terminologies du portail en utilisant un appariement exact.

5.3.2 Structuration des concepts

Dans une ontologie, les concepts sont organisés via des relations taxonomiques et associatives. Cette section s'intéresse à l'identification de ces différentes relations pour structurer les concepts de l'ontologie. Nous présentons dans un premier temps les relations que nous avons récupérées dans l'UMLS puis nous expliquons les actions correctives que nous leur avons appliquées afin d'obtenir un noyau ontologique propre.

[52] https://uts.nlm.nih.gov/home.html

5.3.2.1 Réutilisation des connaissances de l'UMLS

Dans un premier temps, les relations de subsomption (typées *is_a*) directes entre les concepts extraits dans l'étape précédente ont été récupérées à partir du Metathesaurus. Dans le cadre de ce travail, ces relations sont considérées automatiquement comme valides. Cependant, en plus des relations directes, les concepts sont aussi liés par des relations hiérarchiques indirectes. Ainsi, pour tout couple de concepts de la liste initiale, les concepts intermédiaires permettant de les organiser via des relations de subsomption explicites ont aussi été extraits pour une meilleure structuration au sein de l'ontologie. Pour chaque concept de la liste, on parcourt ses ascendants, du plus spécifique au plus général, jusqu'à retrouver d'autres concepts de la liste ou atteindre le nombre de liens maximal fixé (plafond fixé à 7 pour ne pas intégrer les concepts très généraux). Si un concept de la liste est rencontré, le chemin entre le premier concept et ce dernier est récupéré et l'ensemble des concepts se trouvant sur ce chemin (concepts intermédiaires) sont retenus. La figure 11 montre un exemple de concept intermédiaire intégré dans l'ontologie.

En plus de ces concepts intermédiaires, nous avons également sélectionné les descendants (sous-concepts directs et indirects) de neuf concepts spécifiques pour le domaine (Tableau 3) pour enrichir et mieux structurer l'ontologie; une liste des 50 concepts les plus fréquents dans les corpus est présentée à un spécialiste qui sélectionne ces concepts considérés comme très pertinents pour le domaine. Concernant les concepts n'entretenant pas de relations taxonomiques avec les autres, des relations plus générales de l'UMLS (*Child, Parent, Narrower, Broader*) ont été considérées comme des relations taxonomiques candidates, qui ont été validées dans un deuxième temps. En effet, ces relations candidates peuvent concerner différents types de relations sémantiques, telles que *part_of, is_a, mapped_to, contains*. Elles nécessitent ainsi d'être analysées pour garder uniquement les relations pertinentes et éventuellement en distinguer comme étant de différents types au sein de l'ontologie.

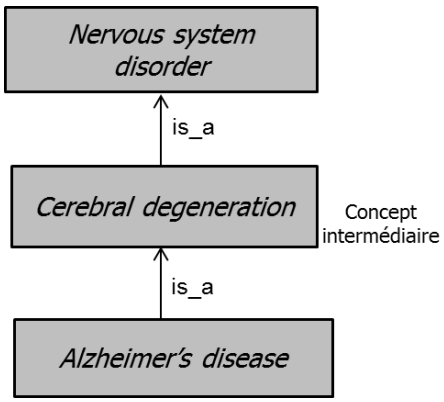


Figure 11 : Exemple de concept intermédiaire intégré automatiquement dans l'ontologie

En plus des relations taxonomiques, nous avons extrait certaines relations transversales typées de manière explicite (par exemple, les relations *may_treat*, *cause_of*, *anatomical_part_of*).

Enfin, nous avons constaté que beaucoup des concepts extraits n'étaient pas rattachés, via des relations taxonomiques, à des concepts généraux et restaient ainsi au premier niveau de l'ontologie. Aussi, pour une meilleure structuration, les types sémantiques du réseau

sémantique catégorisant les concepts ont également été intégrés dans l'ontologie pour regrouper ces concepts. La relation de catégorisation a été remplacée par la relation de subsomption dans notre ontologie.

Tableau 3 : Liste des neuf concepts spécifiques au domaine de la maladie d'Alzheimer

CUI	Terme préféré anglais
C0002395	Alzheimer's disease
C0338656	Impaired cognition
C0011269	Dementia, vascular
C0233794	Memory impairment
C0236642	Pick disease of the brain
C0338451	Frontotemporal dementia
C0752347	Lewy body disease
C0451306	Mini-mental state examination
C1963167	Memory impairment adverse event

5.3.2.2 Actions correctives

Bien que l'UMLS soit une ressource sémantique fournie, elle est limitée par son niveau de spécification superficiel avec des relations vagues comme *Broader*, *Narrower* ou encore *Related Other* (Zweigenbaum, 2004). De plus, la variété des vocabulaires sources et sa large couverture conduisent à des redondances (Peng et al., 2002) et parfois à des incohérences (Cimino et al., 2003; Gu et al., 2004; Bodenreider et al., 2002). Pour pallier certains de ces problèmes, nous avons défini un ensemble de mécanismes et règles inspirés des travaux décrits dans (Chrisment et al., 2008).

Les règles

Règle 1 : Si un couple de concepts est lié par une relation de subsomption, alors toute autre relation existant entre ces deux concepts est redondante. Formellement :

$\forall\, c_1, c_2 \in$ C, si $(c_1, c_2) \in$ H et $r(c_1, c_2) \in$ R, alors la relation $r(c_1, c_2)$ est redondante.

Règle 2 : Si un couple de concepts est lié par différents types de relations, elles-mêmes liées par une relation de subsomption, alors seule la relation sémantique la plus spécifique est retenue. Formellement :

$\forall\, c_1, c_2 \in$ C, si $r_1(c_1, c_2) \in$ R et $r_2(c_1, c_2) \in$ R et $r_1 \subseteq r_2$, alors la relation $r_2(c_1, c_2,)$ est redondante.

Règle 3 : Supposons qu'il existe une relation de subsomption entre les couples de concepts (c_1, c_2) et (c_2, c_3). S'il existe une autre relation de subsomption entre c_1 et c_3, alors cette dernière est redondante. En effet, les relations de subsomption étant transitives, les langages de représentation des ontologies tels que OWL intègrent des mécanismes de déduction qui permettent de déduire automatiquement cette information. Par généralisation, cette règle est définie formellement par :

$\forall\ (c_s, c_t) \in$ H, si $\exists\ \{c_i\}\ (c_s, c_0), (c_0, c_1), \ldots, (c_n, c_t) \in \mathrm{H}^+$, alors la relation de subsomption (c_s, c_t) est **redondante**. La figure 12 montre un exemple de relation de subsomption redondante.

Règle 4 : Au sein de l'ontologie, les relations de subsomption ne doivent pas former un cycle. Formellement H est dit **cyclique** si

$\exists c \in$ C tels que $(c, c) \in \mathrm{H}^+$

Règle 5 : Le « terme préféré » d'un concept ne peut pas apparaitre parmi les termes d'un autre concept.

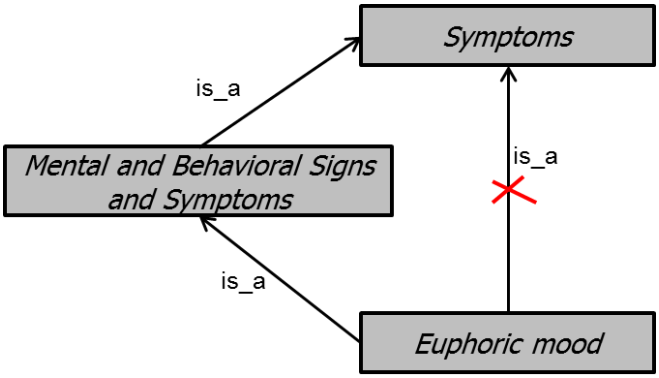


Figure 12 : Exemple de relation de subsomption redondante supprimée par la règle 3

L'application des règles

Des algorithmes ont donc été mis en œuvre pour effectuer les actions correctives nécessaires à la bonne qualité de l'ontologie.

La règle 1 a permis d'éliminer toutes les relations transversales entre deux concepts si une relation de subsomption existait entre eux.

Pour que la règle 2 puisse être appliquée, nous avons établi manuellement une hiérarchie entre les différents types de relations retenus avec l'aide des spécialistes. A cet effet, nous avons étudié des exemples de concepts associés par ces relations.

Pour l'application de la règle 3, le principe est simple. Pour chaque concept c, l'ensemble de ses parents (ses super-concepts) directs, noté $PAR(c)$, est récupéré. Ensuite, pour chacun de ses parents, les chemins menant vers les concepts plus généraux de l'ontologie sont parcourus jusqu'à retrouver un concept de l'ensemble $PAR(c)$ ou atteindre la profondeur maximale de

l'ontologie. Si un concept parent est trouvé dans un chemin, le lien direct entre ce dernier et le concept c est automatiquement supprimé.

En ce qui concerne la règle 4, elle a été définie puisqu'il a été montré que le Metathesaurus de l'UMLS contient des cycles (Bodenreider, 2001). Ces derniers sont étudiés pour supprimer les relations taxonomiques incohérentes. Un algorithme de détection automatique de ces cycles a ainsi été développé. Les cycles identifiés ont ensuite été analysés manuellement pour élaguer les relations incohérentes. La détection des cycles est similaire à la recherche de hiérarchies indirectes. La seule différence est que, pour les cycles, on parcourt les ascendants du concept jusqu'à rencontrer le même concept ou atteindre la longueur maximale du chemin fixée.

La règle 5 a dû être définie puisqu'il a été montré que le Metathesaurus contient des termes polysémiques, rattachés à des concepts différents (Mougin et al., 2009). Pour l'appliquer, nous avons simplement supprimé les occurrences des termes polysémiques apparaissant dans les concepts où ils n'avaient pas le rôle de terme préféré.

5.4 L'enrichissement de l'ontologie

La phase d'enrichissement comprend deux étapes : l'alignement des candidats termes (5.4.1) et l'intégration de nouveaux concepts (5.4.2).

5.4.1 L'alignement des termes

La recherche de concepts associés aux termes repérés dans les corpus met en évidence une large prédominance de l'anglais dans l'UMLS (le pourcentage des termes extraits des corpus et trouvés dans l'UMLS est de 45% pour l'anglais contre 34% pour le français) car les terminologies qui y sont intégrées sont principalement en anglais. Pour compléter les connaissances en français et développer ainsi une ressource bilingue riche du domaine, nous avons proposé une méthode d'alignement de termes (Figure 13) (Drame et al., 2012). Cette méthode s'inscrit dans le cadre des approches d'alignement à partir de corpus parallèles qui sont largement utilisées pour la construction de RTO multilingues (Och et Ney, 2003;) Déjean et al., 2005; Deléger et al., 2009). Elles consistent en la mise en correspondance de termes extraits dans des textes parallèles en utilisant des techniques heuristiques, statistiques ou linguistiques. Une fois le corpus parallèle de textes anglais-français constitué (5.4.1.1), le processus comprend deux phases : l'alignement des phrases (5.4.1.2) et l'alignement des candidats termes (5.4.1.3).

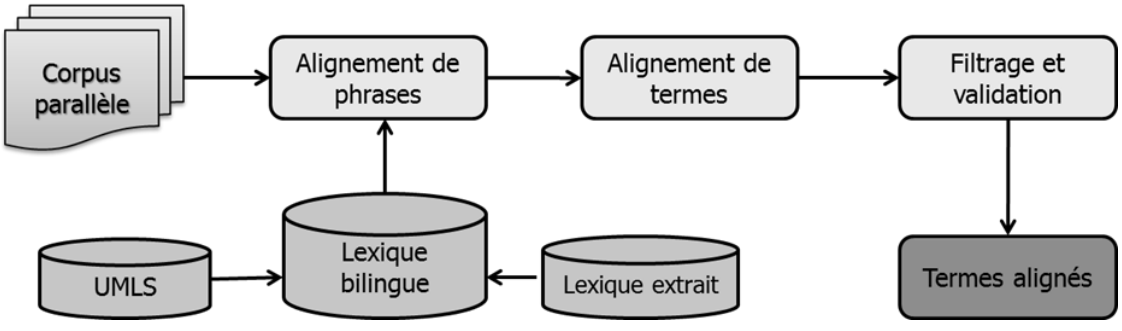


Figure 13 : Méthode d'alignement des termes

5.4.1.1 La constitution du corpus parallèle

A partir de la base de données BiblioDem, nous avons créé un corpus parallèle aligné au niveau des phrases, constitué d'un ensemble de titres d'articles scientifiques en anglais et de leur traduction en français. Cette traduction ayant été réalisée manuellement par des experts du domaine lors de l'analyse des articles, on obtient un corpus parallèle dont la qualité de la traduction est garantie. Ce corpus contient 1 356 paires de titres d'articles anglais - français. Les titres ne contenant qu'une partie des termes pertinents du domaine, la couverture de ce corpus risquait d'être insuffisante. Nous avons ainsi constitué un corpus bilingue (anglais-français) supplémentaire à partir du site Web canadien de la *Société Alzheimer*[53] qui est également dédié à la maladie d'Alzheimer. Dans ce site grand public, la plupart des contenus des pages Web sont exprimés dans les deux langues et sont généralement des traductions les unes des autres. Pour identifier les paires de pages parallèles, leurs URL (Uniform Resource Locator) ont été utilisées. Par exemple, les URL *www.alzheimer.ca/**en**/About-dementia/Alzheimer-s-disease* et *www.alzheimer.ca/**fr**/About-dementia/Alzheimer-s-disease* correspondent à la même page dans les deux langues. Ainsi, une collection de 705 paires de documents parallèles a été obtenue après nettoyage et filtrage d'un corpus initial contenant toutes les pages du site Web. Ensuite, le contenu textuel des pages Web a été extrait et nettoyé avec l'analyseur *Parser HTML*[54]. C'est un outil simple qui analyse, extrait et filtre le contenu des pages Web. Les titres des différents documents ont été alignés directement car nous avons considéré que le titre d'un document anglais correspondait au titre du document français correspondant. Au total, 8 766 paires de phrases anglais-français ont été créées à partir du site de la *Société Alzheimer*. Notre corpus parallèle contient donc au final 10 122 paires de phrases anglais-français.

5.4.1.2 L'alignement des phrases

Pour aligner les contenus textuels des documents parallèles, les corpus ont d'abord été découpés en phrases en utilisant l'identificateur de phrases intégré dans l'outil *OpenNLP*[55]. Ensuite, une méthode d'alignement de phrases a été utilisée pour les mettre en correspondance. Pour aligner des phrases à partir de documents parallèles, les approches utilisées dans la littérature sont diverses : certaines sont basées sur la longueur des phrases (Gale et Church, 1991) tandis que d'autres exploitent un lexique bilingue et/ou la similarité morphologique entre les mots de langues différentes (Melamed, 1999). Dans ce travail, nous utilisons l'outil *GMA*[56] (Geometrical Mapping and Alignment), qui est basé sur la méthode proposée dans (Melamed, 1999). Il combine différentes techniques d'appariement qui exploitent les cognats orthographiques et/ou la similarité entre tokens basée sur la plus longue séquence de caractères commune (*Longest Common Subsequence*) entre eux. Cet outil utilise également un lexique bilingue.

[53] http://www.alzheimer.ca/
[54] http://htmlparser.sourceforge.net
[55] http://opennlp.apache.org/
[56] http://nlp.cs.nyu.edu/GMA/

Pour cela, un lexique bilingue extrait à partir de l'UMLS, en considérant les termes dans les deux langues ayant le même CUI comme synonymes, a été fusionné avec celui généré en utilisant seulement les titres des articles de BiblioDem (Drame et al., 2012).

Le lexique bilingue résultant et le corpus parallèle forment les entrées de l'outil GMA. Ainsi, à partir de ces textes parallèles, GMA retourne un corpus bilingue constitué de 10 122 phrases alignées. Dans les résultats de cet outil d'alignement, une phrase du texte source peut correspondre à une ou, plus rarement, à plusieurs phrases du texte cible.

5.4.1.3 L'alignement des termes

Pour aligner les termes repérés dans le texte source avec leurs équivalents dans le texte traduit, nous avons proposé, dans un premier temps, une méthode combinant deux techniques (Drame et al., 2012) : une technique heuristique basée sur le calcul de score d'association et une méthode linguistique fondée sur la similarité morphologique.

La technique heuristique est basée sur les cooccurrences entre un terme source et un terme cible pour estimer leur degré d'association. On se base sur l'hypothèse suivante : *les termes apparaissant dans des portions de textes parallèles sont susceptibles d'être en relation de traduction*. Le score d'association choisi, l'indice de Jaccard (Jaccard, 1912), est couramment utilisé dans la littérature et il est défini pour tout couple de termes t_s et t_c par :

$$jaccard(t_s, t_c) = \frac{coocc(t_s, t_c)}{freq(t_s) + freq(t_c) - coocc(t_s, t_c)}$$

avec $coocc(t_s, t_c)$, le nombre de cooccurrences dans des phrases parallèles de t_s et t_c et $freq(t)$, le nombre d'occurrences du terme t.

En plus d'être simple, cette mesure permet de déterminer les couples de termes les plus fortement associés dans un corpus aligné. Pour chaque terme de la langue source, ses potentiels correspondants sont les termes de la langue cible dont le score de Jaccard dépasse un certain seuil. Pour vérifier la réciprocité de l'équivalence entre termes, nous avons réalisé l'appariement dans les deux sens en considérant d'une part l'anglais comme langue source et le français comme langue cible et inversement.

La technique linguistique se fonde sur le fait que les termes médicaux (pour les langues française et anglaise) possèdent de nombreuses racines gréco-latines. Ainsi, nous avons choisi d'utiliser la distance de *Levenshtein* (Levenshtein, 1966) normalisée qui permet de repérer les couples de termes identiques ou similaires morphologiquement dans les deux langues, tels que : *atrophie cérébrale* et *cerebral atrophy*. Cette mesure s'est montrée efficace et a enregistré de bonnes performances pour le calcul de la similarité entre deux chaines de caractères (Okuda et al., 1976; Soukoreff et MacKenzie, 2001; Haldar et Mukhopadhyay, 2011).

Appliquée sur un corpus parallèle constitué seulement des titres des articles en anglais et leur traduction en français, cette méthode avait donné de bons résultats avec une précision de 73%. Les performances de la méthode se sont révélées moins bonnes sur ce corpus étendu avec le

site Web de la Société Alzheimer (cf section 5.4.1.1), plus large mais bruité. Nous l'avons ainsi combinée à l'aligneur statistique de *Moses* (Koehn et al., 2007) (cf section 4.3.2) pour améliorer les résultats de l'alignement. Cet outil établit des correspondances entre les mots et utilise les cooccurrences des mots et des séquences de mots (phrases) pour fournir des alignements entre les groupes de mots. En plus d'être l'un des outils de traduction automatique les plus utilisés, il a montré de bonnes performances dans des expérimentations réalisées sur des collections standards (Lardilleux et al., 2012; Ren et al., 2009). Le principe de Moses est le suivant : les corpus sont segmentés en mots, normalisés (chaque mot est converti en sa forme la plus fréquente dans le corpus) et nettoyés en supprimant les phrases très longues et celles n'ayant pas de correspondant. Ensuite, une table de traduction (fournissant les alignements entre les segments de la langue source et ceux de la langue cible) est automatiquement générée à partir de ces corpus prétraités. Enfin, les alignements sont filtrés pour sélectionner seulement les candidats termes retenus dans la première étape. En plus, seules les paires de termes dont les probabilités de traduction dépassent un seuil minimal (choisi empiriquement en se basant sur les expérimentations décrites dans (Lardilleux et al., 2012)), sont sélectionnées pour optimiser à la fois la précision et le rappel. Une fois les termes appariés, ceux-ci ont été présentés aux experts du domaine pour validation. Lorsque l'appariement était jugé correct, le terme traduit a été intégré automatiquement au sein de l'ontologie comme synonyme du concept approprié.

5.4.2 L'intégration de nouveaux concepts

La deuxième étape de la phase d'enrichissement concerne l'intégration de nouveaux concepts dans l'ontologie. Pour améliorer la couverture des connaissances, des concepts spécifiques du domaine qui ne sont pas contenus dans les vocabulaires sources de l'UMLS ont dû être intégrés. Les candidats termes retenus lors de la première étape mais n'ayant pas été trouvés dans l'UMLS constituaient de bons candidats pour étendre l'ontologie. Les candidats termes des deux langues correspondant à des variants lexicaux ou des synonymes (obtenus grâce aux méthodes d'alignement) ont d'abord été regroupés dans un même concept automatiquement si la traduction avait été validée. Par exemple, le terme *Cognitive test scores* et *Scores aux tests cognitifs* sont regroupés dans le même concept (AD000087, identifiant créé lors de son intégration au sein de notre ontologie). Dans le cas où un terme à ajouter n'a pas de synonyme dans une langue, il est aussi conceptualisé. Par la suite, les relations de dépendance en tête issues de l'analyse syntaxique des corpus ont permis de placer les nouveaux concepts au sein de l'ontologie. Comme dit précédemment (section 4.3.1), après une analyse du corpus, Syntex fournit un réseau terminologique où chaque candidat terme est lié à l'ensemble de ses *descendants en tête* (c'est-à-dire les candidats termes l'ayant comme tête). Nous avons donc exploité ces liens syntaxiques pour établir des relations taxonomiques entre les nouveaux concepts et ceux de l'ontologie. Cette technique consistant à exploiter la structure interne des termes multi-mots a été largement explorée dans la construction d'ontologies à partir de textes (Buitelaar et al., 2004; Velardi et al., 2006). Le principe est le suivant : si un terme candidat t_1 a pour tête un autre candidat t_2, alors le concept associé à t_1 est défini comme un enfant du concept correspondant à t_2. Formellement :

$$si\ \exists\ t_1, t_2\ \in T\ avec\ t_2\ t\hat{e}te\ de\ t_1\ \text{alors}\ \big(F(t_1), F(t_2)\big) \in H$$

Par exemple, le candidat terme *Severe Alzheimer disease* associé au concept *AD000390* étant un descendant en tête du candidat terme *Alzheimer disease* dénotant le concept *C0002395*, on en déduit que *AD000390* est un sous-concept de *C0002395*. La figure 14 donne d'autres exemples de relations de dépendance en tête. Dans le cas où le concept parent n'est pas intégré dans l'ontologie, celui-ci est placé au premier niveau de l'ontologie et l'enfant est ensuite connecté directement à ce dernier.

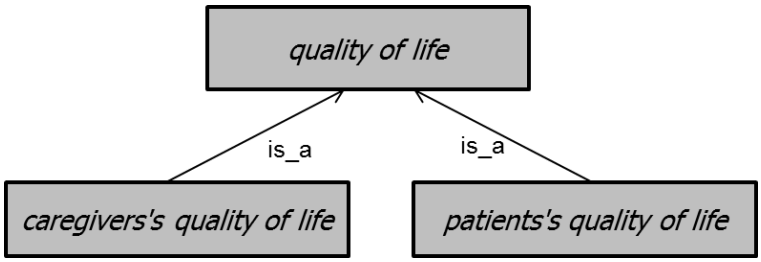


Figure 14 : Exemples de relations de dépendance en tête

Une fois les mécanismes de construction de l'ontologie appliqués, il a fallu valider le contenu de l'ontologie. Une fois cette étape majeure effectuée, l'ontologie a été décrite de manière formelle.

5.5 La validation et la formalisation de l'ontologie

5.5.1 Validation

L'ensemble des résultats a été validé par deux spécialistes de la maladie d'Alzheimer : une interne de santé publique et une doctorante en épidémiologie, supervisées par un professeur en Neurologie, tous membres de l'équipe *Epidémiologie et Neuropsychologie du Vieillissement Cérébral*. Le processus de validation a été subdivisé en différentes étapes pour le simplifier. Ainsi, les experts ont été sollicités pour valider les éléments suivants :

- Les concepts retrouvés dans l'UMLS. Pour cela, les concepts leur ont été fournis avec quelques termes associés et leurs contextes d'apparition dans les corpus. Ces informations leur ont été présentées sous forme de tableau Excel. Ceci a permis aux validateurs de comprendre les contextes de génération des concepts mais aussi de les manipuler facilement.

- Les relations taxonomiques candidates issues des relations de dépendance de Syntex ont été analysées par les spécialistes pour déterminer celles qui représentent de vraies relations de subsomption.

- Les différents types de relations non taxonomiques extraits de l'UMLS ont été aussi analysés pour filtrer ceux qui sont pertinents pour notre application visée.

- Les alignements. Les paires de termes obtenues grâce à l'alignement et correspondant aux mêmes concepts dans l'UMLS ont été validées automatiquement. Si l'un des termes n'avait pas été retrouvé dans l'UMLS, les paires ont été présentées aux spécialistes.

- Les nouveaux concepts qui sont parfois des concepts spécifiques du domaine. Les spécialistes ont d'abord vérifié si les concepts sont pertinents pour le domaine. Ils ont ensuite dû valider les liens taxinomiques les rattachant aux autres concepts du noyau ontologique.

Les deux experts ont examiné les résultats de ces différentes étapes. Lorsque leurs validations différaient, ils devaient en discuter pour parvenir à un consensus.

5.5.2 Formalisation de l'ontologie

Cette étape finale consiste à décrire l'ontologie résultant des étapes précédentes dans un langage formel et expressif. Pour cela, le langage de référence OWL a été utilisé. Celui-ci permet de représenter certaines propriétés complexes des relations, telles que le fait d'être inverses (comme par exemple, les relations *part_of* et *has_part*), la symétrie et la transitivité. Représenter l'ontologie dans ce langage facilite également l'interopérabilité de cette ressource avec les autres qui sont, pour la plupart, décrites dans ce langage. Nous avons par ailleurs utilisé le langage SKOS puisqu'il est particulièrement approprié pour supporter le multilinguisme ; aspect majeur pour représenter notre ontologie bilingue. Ainsi, chaque concept de l'ontologie est représenté par une classe OWL ayant les propriétés suivantes :

- un *URI* qui représente son identifiant unique ;
- un terme préféré (*skos:prefLabel*) en anglais et éventuellement un terme préféré en français ;
- des termes synonymes (*skos:altLabel*) en anglais et en français ;
- éventuellement des définitions (*skos:définition*) dans les deux langues.

Les relations transversales structurant les concepts sont définies par des propriétés OWL du type ObjectProperty. La figure 15 montre le modèle conceptuel utilisé pour représenter les entités de l'ontologie.

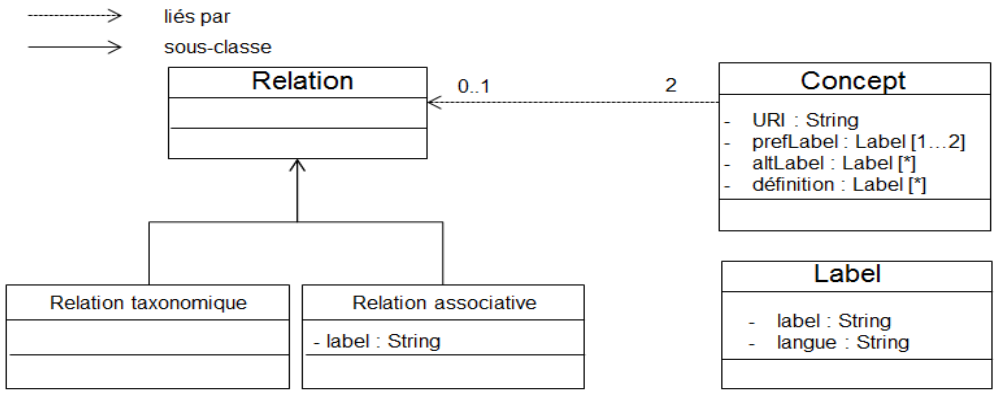


Figure 15 : Modèle de représentation des entités de l'ontologie

6 Résultats

Dans cette partie, nous présentons les résultats de l'application de notre méthodologie sur le domaine spécifique de la maladie d'Alzheimer.

Tableau 4 : Les 10 syntagmes nominaux les plus fréquents dans les corpus anglais et français

Syntagme nominal en anglais	Fréquence dans le corpus anglais	Syntagme nominal en français	Fréquence dans le corpus français
Alzheimer's disease	1736	maladie d'Alzheimer	1452
cognitive impairment	764	déclin cognitif	586
mild impairment	437	sujets âgés	428
mild cognitive impairment	419	facteurs de risque	311
cognitive decline	375	performances cognitives	298
cognitive function	273	fonctions cognitives	298
risks factors	241	troubles cognitifs	293
confidence interval	221	risque de démence	275
patients with disease	186	niveau d'études	269
risk of dementia	183	personnes âgées	257

6.1 L'extraction des candidats termes

Un ensemble de 49 390 syntagmes nominaux et 8 844 noms ont été extraits à partir du corpus anglais. De manière similaire, dans le corpus français, 69 505 syntagmes nominaux et 11 688 noms ont été extraits. Le tableau 4 présente les dix syntagmes nominaux les plus fréquents dans les corpus anglais et français. Après filtrage, 2 916 candidats termes anglais ont été retenus, correspondant à 1083 candidats termes simples (c'est-à-dire composés d'un seul mot) et 1833 candidats termes complexes (c'est-à-dire composés de plusieurs mots). Dans le corpus français, 3 152 candidats termes ont été trouvés : 1 196 candidats termes simples et 1 956 candidats termes complexes. A titre d'exemple, les candidats termes *prior research* et *optimal mechanisms* ont été élagués car leur fréquence dans le corpus était inférieure à sept.

6.2 La construction du noyau ontologique

Dans la phase de conceptualisation, une bonne partie des candidats termes simples retenus dans l'étape précédente ont été retrouvés dans le Metathesaurus de l'UMLS (65%) tandis que pour les candidats termes complexes, un peu moins du tiers (32%) est aligné avec au moins une entrée de cette ressource. Ceci peut s'expliquer par le fait que les syntagmes nominaux sont plus spécifiques au domaine et, par conséquent, sont moins systématiquement présents dans cette ressource médicale plus générale. Par exemple, les termes *episodic memory impairment* et *severe Alzheimer disease* ne sont pas retrouvés dans l'UMLS et on notera qu'ils sont pourtant pertinents pour décrire le domaine d'intérêt. Comme prévu, nous avons également observé une meilleure couverture des termes anglais : 45% contre 34% pour les candidats termes français en combinant les résultats de l'UMLS et de CISMeF.

De l'alignement des candidats termes aux entrées du Metathesaurus résulte un ensemble de 3 871 concepts. Après validation par les experts, 2 421 concepts ont été retenus (62%). Par exemple, les concepts *Physical activity* (C0026606), *Physiological stress* (C0449430) et *Risk factors* (C0035648) ont été jugés valides tandis que *Scientific control* (C1882979), *Science of anatomy* (C0002808) et *Number of patients* (C2360800) ont été invalidés.

En plus de ces concepts, 2 905 concepts supplémentaires (intermédiaires et spécifiques) ont été intégrés dans le but de mieux structurer l'ontologie. Par exemple, avec les concepts *Dementia* (C0497327) et *Mental disorders* (C0004936) initialement trouvés, le concept intermédiaire *Organic psychiatric disorders* (C2013984) a été ajouté.

L'ensemble de ces 5 326 concepts sont associés via 7 499 relations taxonomiques et 10 889 relations transversales. En effet, à partir de 8 125 relations taxonomiques initialement extraites de l'UMLS, 903 relations redondantes ont été supprimées. Deux cycles ont également été identifiés et corrigés. Par exemple, pour le cycle *psychometric projective* (C2143019) → *psychometrics* (C0033920) → *psychological test* (C0033905) → *psychometric* (C2143019), la relation taxonomique entre les concepts *psychological test* (C0033905) et *psychometric* (C2143019) a été supprimée. De plus, la hiérarchie de l'ontologie a été enrichie de 279 relations taxonomiques générées à l'aide des dépendances en tête de Syntex. Par exemple, *Severe Alzheimer disease* est un sous-concept de *Alzheimer disease*. En outre, des relations transversales de différents types ont été conservées. Ainsi, parmi les 221 types de relations liant les concepts extraits des corpus, 178 ont été jugés pertinents par les experts du domaine. Parmi les types de relations retenus, 82 sont l'inverse d'une autre. Le tableau 5 montre les six types de relations les plus fréquents dans l'ontologie avec des exemples de concepts qu'elles lient.

6.3 L'enrichissement de l'ontologie

Une table de traduction de 492 556 lignes a été générée avec *Moses* à partir du corpus parallèle présenté en section 4.3.2. En filtrant ces alignements sur les candidats termes extraits et retenus des corpus, augmentés de leurs synonymes extraits du Metathesaurus, on obtient un ensemble de 1959 paires de candidats termes alignés. Les résultats de l'alignement varient en

fonction du seuil de probabilité fixé pour le filtrage. Plus la probabilité est élevée, plus les candidats termes sont susceptibles d'être en relation de traduction. Ainsi, plus le seuil est élevé, plus la précision augmente tandis que le rappel décroit et vice versa.

Tableau 5 : Les types de relations transversales les plus fréquents avec leur nombre d'occurrences dans l'ontologie et des exemples de concepts qu'elles lient

Relation	Occurrences	Concept source	Concept cible
clinically_associated_with	1684	Hypertensive disease	Dementia
has_finding_site	443	Presenile dementia	Brain
contraindicated_drug	374	Physostigmine	Cardiovascular diseases
gene_encodes_gene_product	363	MAPT gene	Microtubule-associated protein tau
has_associated_morphology	292	Down syndrome	Congenital abnormality
disease_has_associated_anatomic_site	278	Cerebral infarction	Cardiovascular system

Tableau 6 : Résultats de l'alignement des termes anglais-français en fonction des seuils de probabilité de traduction fixés

Seuil de probabilité minimal	Nombre total d'alignements	Nombre d'alignements corrects	Précision
0,5	1013	752	74,2%
0,6	727	586	80,6%

Le tableau 6 illustre ce comportement en considérant deux seuils différents. Ainsi, en fixant des seuils de 0,5 et 0,6, respectivement 1013 et 727 paires de candidats termes alignés ont été obtenues. Même si la précision était moindre pour le seuil à 0,5, nous avons choisi d'utiliser celui-ci afin de détecter un plus grand nombre de paires de termes alignés valides. On aurait également pu choisir un seuil plus bas mais ceci aurait nécessité plus d'efforts de validation avec des résultats moins précis.

Puisque nous ne disposions pas d'alignements validés constituant une référence, le rappel n'a pas pu être évalué. Le tableau 7 donne des exemples de couples de termes alignés grâce à *Moses* avec leurs probabilités conditionnelles de traduction correspondantes. Avec ce traducteur statistique, toute séquence de mots est considérée et est mise en relation avec tous ses potentiels correspondants. Par exemple, l'expression en français *aphasie primaire progressive* (notée f) peut être alignée aux expressions anglaises (notées e_i) *primary*

*progresssive aphasi*a, *of progressive aphasia* et *progressive aphasia* avec respectivement comme probabilité que f soit une traduction de e_i (i.e., $P(f|e_i)$) : 0,09, 0,05 et 0,33. Inversement, les valeurs de $P(e_i|f)$ sont respectivement de : 1, 0,5 et 0,5. Notons que les traductions dans le corpus parallèle n'étant pas totalement symétriques (à cause des problèmes de synonymie ou de traductions partielles ou erronées), l'écart entre les probabilités conditionnelles varie considérablement. Ainsi, après le premier filtrage sur les listes de candidats termes, un deuxième filtrage permet de sélectionner seulement les alignements dont une des probabilités conditionnelles dépasse le seuil minimal de probabilité de traduction fixé. Les résultats obtenus avec le seuil minimal fixé à 0,5 ont été combinés à ceux obtenus dans (Drame et al., 2012), résultant finalement en un total de 1527 alignements. Notons qu'une bonne partie des termes alignés existaient déjà dans l'UMLS ou dans la SNOMED 3.5. Ainsi, l'ontologie a été finalement enrichie par l'intégration de 608 synonymes français supplémentaires. Par exemple, les synonymes suivants ont été ajoutés à l'ontologie : *déclin cognitif rapide* (*rapid cognitive decli*ne) et *troubles comportementaux* (*behavioral disturbances*).

Tableau 7 : Exemples de paires de termes alignés avec Moses avec les probabilités conditionnelles de traduction

Terme français f	Terme anglais e	P(f|e)	P(e|f)
Aphasie primaire progressive	Primary progressive aphasia	0,09	1
Trouble cognitif léger	Mild cognitive impairment	0,01	1
Activité physique	Physical activity	0,43	0,77
Cause de décès	Cause of death	1	1
Activation microgliale	Microglial activation	1	1
Antécédents familiaux	Family history	0,67	0,29
Troubles neurologiques	Neurological disorders	0,75	0,6
Comportements agressifs	Aggressive behavior	1	0,5
Troubles psychologiques	Psychological symptoms	1	0,5
Facteur de risque génétique	Genetic risk factor	1	1

Par ailleurs, 439 nouveaux concepts (hors UMLS) ont été intégrés pour étendre la première mouture de l'ontologie. Rappelons que ces concepts ont été validés manuellement par les experts mais qu'ils sont ensuite placés de manière automatique dans l'ontologie lorsqu'une relation de dépendance en tête était fournie par Syntex. Par exemple, le concept *Episodic memory impairment* a été ajouté comme un sous-concept de *Memory impairment* à l'aide des dépendances syntaxiques.

Le temps passé pour la validation de chaque étape est présenté dans le tableau 8.

L'approche proposée a donc permis de construire une ontologie partiellement bilingue spécifique à la maladie d'Alzheimer, OntoAD, contenant 5 765 concepts dont 3 283 (56,9%) ont des synonymes en français. Ces concepts sont structurés via 7 499 relations taxonomiques

et 10 889 relations transversales. La figure 16 présente une portion de l'ontologie résultante qui est disponible en ligne à l'adresse suivante : http://lesim.isped.u-bordeaux2.fr/SemBiP/ressources/OntoAD.owl.

Tableau 8 : Temps de validation pour chaque étape

Tâches	Temps de validation (heures/personne)
Concepts extraits (3871)	25
Relation taxonomiques candidates (499)	5
Types de relations transversales (221)	1
Alignements (2069)	15
Nouveaux concepts (439)	15
Total	61

En termes de connaissances, l'ontologie regroupe des concepts spécifiques du domaine (*Démence*, *Déclin cognitif léger*, *Maladie de Parkinson*, etc.) mais aussi des concepts plus génériques jugés néanmoins pertinents, tels que des concepts méthodologiques (*Prévalence statistique*, *Etude de cohorte*, etc.) qui sont couramment utilisés en épidémiologie. L'ontologie contient aussi des concepts médicaux d'ordre général qui sont utiles pour la structuration des concepts spécifiques (*Human body structure*, *Blood pressure*, etc.) ou qui permettent de décrire des indicateurs propres à la maladie d'Alzheimer (*Facteurs de risques*, *Niveau d'études*, etc.).

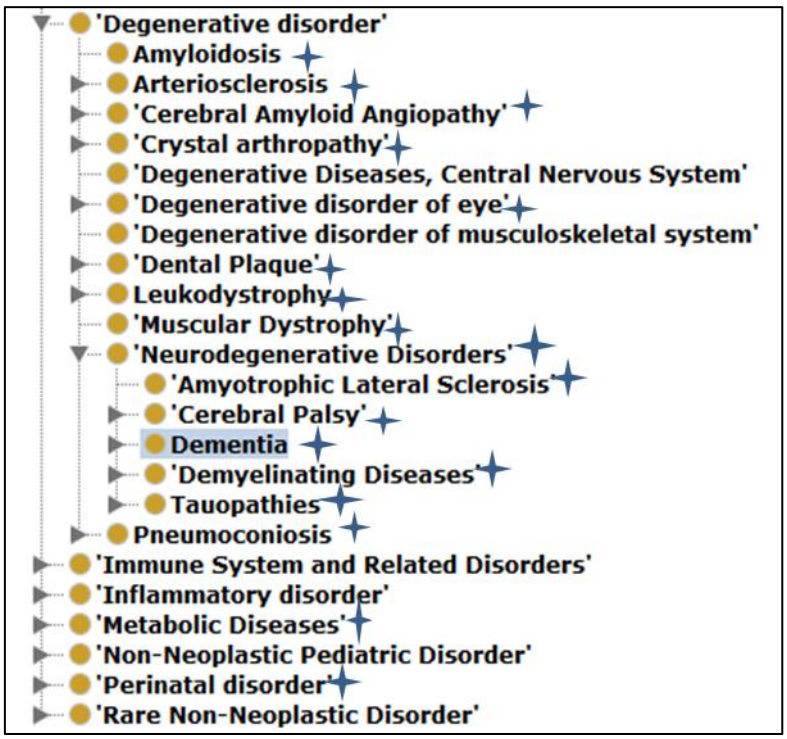


Figure 16 : Visualisation d'une portion de l'ontologie dans le logiciel Protégé. Le signe « étoile » indique que les concepts ont au moins un synonyme en français

7 Discussion

L'application de l'approche proposée dans ce travail au domaine de la maladie d'Alzheimer a permis d'extraire, semi-automatiquement, une première mouture d'une ontologie partiellement bilingue. Vu le nombre élevé de concepts validés, nous considérons que l'utilisation de corpus spécifiques du domaine (extraits de BiblioDem) a permis de générer un modèle de connaissances qui couvre différents aspects liés à la maladie d'Alzheimer. Nous verrons par la suite (chapitre 4) qu'une telle ontologie, extraite à partir des documents d'une base bibliographique, est particulièrement adaptée et utile pour supporter la RI conceptuelle sur ces documents.

Au-delà des concepts, de nombreuses relations ont été extraites pour structurer l'ontologie. Bien qu'elles aient été jugées pertinentes par des spécialistes du domaine, certaines relations (par exemple, *clinically_associated_with*) sont vagues et pourtant très fréquentes dans l'ontologie, ce qui n'apporte pas un sens assez précis du lien existant entre les concepts. Une spécification de ce type de relations serait nécessaire pour mieux expliciter le contenu de l'ontologie.

Par ailleurs, l'utilisation des techniques d'alignement a permis de trouver des synonymes français à des concepts issus de l'UMLS qui n'en possédaient pas. Grâce à notre méthode d'appariement de termes anglais-français, le pourcentage des concepts de notre ontologie possédant des synonymes français est ainsi passé de 46 % à 57 %. Toutefois, une bonne partie des concepts ne sont pas toujours associés à des termes français, probablement parce que notre corpus parallèle ne couvre pas suffisamment les connaissances du domaine de la maladie d'Alzheimer. Notons cependant que si l'on considère seulement les termes extraits à partir de nos corpus (sans prendre en compte les concepts intermédiaires), 71 % sont associés à des synonymes français.

Notre méthodologie peut aussi servir à la construction et à l'enrichissement d'ontologies pour tout sous-domaine de la biomédecine. Cette méthodologie peut également être appliquée à d'autres domaines en tenant compte de leur spécificité et de la disponibilité des ressources sémantiques. Cet aspect est développé de manière plus conséquente dans le dernier chapitre.

Ce travail soulève toutefois de nouvelles questions. Etant donné que tous les termes retrouvés dans les corpus ne peuvent pas être intégrés dans une ontologie (trop grand nombre, non pertinence de certains), un filtrage permettant de sélectionner les termes pertinents est nécessaire. Pour cela, nous nous sommes basés sur leur fréquence (nous avons choisi ceux qui apparaissaient au moins sept fois dans le corpus) en considérant les termes les plus fréquents dans nos corpus comme les plus représentatifs. Ce choix peut cependant se révéler problématique ; certains termes très fréquents ne sont pas pertinents (par exemple, le terme *Outcome Measures* qui apparait 158 fois dans le corpus n'est pas pertinent) tandis que des termes rares peuvent s'avérer pertinents (par exemple, *atrophy of the entorhinal cortex* et *episodic memory testing* apparaissent une fois dans le corpus, *Alzheimer disease risk* et *poststroke dementia* apparaissent six fois et ils sont pertinents). Le développement de mesures statistiques (ou linguistiques) permettant de distinguer efficacement ceux sont pertinents des non pertinents parmi une liste de termes extraits dans un corpus reste une question ouverte.

Un autre challenge concerne la couverture du domaine par les ressources sémantiques exploitées. Dans notre cas, l'UMLS a été utilisée. Cette dernière, bien que couvrant largement le domaine biomédical, reste limitée pour des sous-domaines spécifiques comme la maladie d'Alzheimer. Pour combler cette limite, nous avons proposé une étape d'enrichissement où de nouvelles entités non contenues dans cette ressource sont intégrées au sein de l'ontologie. Une autre piste est la mise en place de techniques pouvant gérer l'évolution de l'ontologie. Pour cela, les traces d'utilisation (logs utilisateur) du portail sémantique basé sur OntoAD pourraient être exploitées afin d'intégrer de nouvelles entités dans l'ontologie. Par ailleurs, l'exploration de techniques de « crowdsourcing », auxquelles s'intéressent de plus en plus les chercheurs en ingénierie des connaissances, parait une perspective intéressante pour enrichir et améliorer le contenu de l'ontologie.

Nous notons aussi que, dans la version actuelle de l'ontologie, beaucoup de nouveaux concepts sont rattachés directement à la racine. Ceci pose des problèmes de structuration et l'utilisation de techniques allant au-delà des dépendances en tête (telles que les patrons morphosyntaxiques) pour établir des relations taxonomiques entre les concepts pourrait être intéressante.

8 Conclusion

Dans ce chapitre, nous avons présenté une approche de construction d'ontologies basée sur l'exploitation de corpus textuels et la réutilisation de ressources sémantiques existantes. C'est une approche simple (même si elle s'appuie sur des techniques complexes) qui a permis de générer une ontologie partiellement bilingue de la maladie d'Alzheimer de manière semi-automatique. Après avoir constitué les corpus de texte, ceux-ci ont été analysés avec l'outil de TAL Syntex pour extraire les candidats termes du domaine. L'UMLS a ensuite été utilisé pour regrouper les candidats termes extraits en concepts et les structurer grâce à des relations taxonomiques et transversales. Toutefois, malgré sa richesse, l'UMLS reste une ressource imparfaite contenant notamment des redondances et des incohérences. Ainsi, des actions correctives ont été proposées pour traiter ces différents problèmes. Par ailleurs, notre approche intègre une phase d'enrichissement comprenant l'alignement de termes de langues différentes en utilisant des techniques de traduction automatique et l'intégration automatique de nouveaux concepts au sein de l'ontologie en exploitant les dépendances syntaxiques entre les termes associés à ces concepts. L'ensemble des connaissances ont été validées par des spécialistes du domaine.

L'ontologie développée est actuellement utilisée pour supporter un portail pour la RI et la navigation sémantiques sur une collection de documents scientifiques dédiée à la maladie d'Alzheimer. Nous montrons dans le chapitre suivant que l'exploitation de l'ontologie en RI est particulièrement intéressante car celle-ci offre de nombreuses fonctionnalités : indexer les documents de manière conceptuelle, guider l'utilisateur dans la formulation de sa requête et enrichir l'ontologie en exploitant les termes présents dans les requêtes mais qui n'apparaissent pas dans l'ontologie elle-même.

Chapitre 4: Indexation et recherche d'information biomédicale basées sur une ressource termino-ontologique

1 Introduction

Pour faire face aux problèmes soulevés dans la RI classique (ambiguïté, disparité des termes, etc.), mentionnés dans le chapitre 2, de nombreux travaux se sont intéressés à la prise en compte de la sémantique soit inhérente aux données traitées, soit fournie par une ressource externe. Ainsi, certaines recherches portent sur l'exploitation de la sémantique latente contenue dans les documents en utilisant des techniques statistiques (Dumais, 1994; Letsche et Berry, 1997) tandis que d'autres s'appuient sur des ressources sémantiques explicites (Egozi et al., 2011; Baziz et al., 2005; Hliaoutakis et al., 2006; Fernández et al., 2011). Dans la seconde approche, les ressources utilisées vont de ressources générales, telles que WordNet (Fellbaum, 1998) et Wikipédia, à des ressources spécifiques à un domaine particulier, comme le thésaurus MeSH ou l'UMLS (Bodenreider, 2004), largement exploités dans la RI biomédicale. Bien que l'utilisation de telles ressources permette d'améliorer les performances (Fernández et al., 2011), elle soulève de nouveaux challenges tels que la disponibilité de ressources adaptées pour le domaine d'application, l'identification des descripteurs sémantiques dans les documents (Suominen et al., 2013), la sélection automatique, parmi ces descripteurs, des plus pertinents pour représenter les documents ou encore l'omission de concepts pertinents dans les ressources (Bhagdev et al., 2008).

Dans ce cadre, nous proposons une approche qui explore le potentiel des ontologies pour améliorer les performances en RI en s'intéressant aux questions soulevées ci-dessus. Et ce avec pour objectif de mettre en œuvre un portail sémantique de RI.

Afin de mieux cerner l'objectif du travail qui a été mené dans le cadre de ce chapitre, nous revenons rapidement sur la notion de portail sémantique et les différentes fonctionnalités qu'ils peuvent offrir.

Les portails sémantiques sont des portails Web basés principalement sur les technologies du Web sémantique (Maedche et al., 2001; Contreras et al., 2004). Leur objectif est de faciliter le partage et l'échange d'informations à une large communauté d'utilisateurs (Zhang et al., 2005). Pour cela, ils offrent différentes fonctionnalités dont une des plus communes est un service de RI sémantique. Par exemple, le portail *OntoFrame S*3 est dédié essentiellement à la RI académique (Lee et al., 2010). Son principe est de s'appuyer sur une ontologie[57] modélisant les agents, tels que les chercheurs et les institutions, leurs réalisations telles que leurs publications et leurs rapports, leurs affiliations, leurs domaines de recherche, etc. pour permettre aux chercheurs d'accéder efficacement à des informations pertinentes dans le cadre de leur recherche scientifique. Les portails sémantiques sont aussi conçus dans le but de permettre aux utilisateurs de naviguer et de visualiser de manière conviviale des données

[57] http://isrl.kisti.re.kr/ontologies/ReferenceOntology1_0.owl

(Ding et al., 2010). Ils supportent également la RI par facette où des facettes intuitives sont dérivées d'ontologies pour aider l'utilisateur à formuler ses requêtes (Suominen et al., 2007).

Dans le domaine de la santé, des portails sémantiques spécifiques ont été développés pour faciliter l'accès à l'information (McGuinness et al., 2012). GoPubMed (Doms et Schroeder, 2005), qui permet d'exploiter les publications stockées dans PubMed en se basant sur l'ontologie GO, est un exemple de portail couramment utilisé dans le domaine. Il fournit une fonctionnalité avancée de RI intégrant des modules d'auto-complétion et de visualisation intuitive des résultats. Il permet également aux utilisateurs de naviguer dans la hiérarchie de l'ontologie pour explorer les documents indexés par les concepts.

Nous retiendrons ainsi qu'un portail sémantique offre généralement un service d'indexation et de recherche sémantique, la mise à jour de la collection de documents gérés, la gestion de la ressource sémantique utilisée pour ces tâches ou encore l'administration des utilisateurs. Dans la suite de ce chapitre, nous avons tenté de répondre aux questionnements soulevés pour la mise en place de certaines de ces fonctionnalités.

Pour cela, nous avons développé une méthode d'indexation sémantique qui comprend deux étapes : 1) **une phase de repérage de descripteurs sémantiques** dans des corpus en privilégiant les plus spécifiques. Dans notre cas, il s'agit de concepts issus d'une RTO, 2) **une méthode de désambiguïsation de concepts basée sur la similarité sémantique**. L'identification des concepts permettant de représenter un document est une phase préalable à l'indexation conceptuelle. Ainsi, nous proposons deux méthodes de repérage de concepts dans des textes biomédicaux. Ces dernières sont positionnées parmi les approches d'extraction de concepts à base de dictionnaires. Les éventuels concepts ambigus sont ensuite désambiguïsés en exploitant leur degré de similarité par rapport aux autres concepts de leur contexte. Pour la pondération, la mesure TF.IDF (term frequency – inverse document frequency) a été adaptée aux concepts. Ces propositions ont été implémentées et validées dans le cadre de la mise en œuvre du portail SemBiP.

Nous présentons également, pour la phase de recherche, l'implémentation d'**une approche d'expansion de requêtes** dans le cadre de SemBiP, reposant sur la similarité sémantique entre les concepts. Enfin, puisque les modèles de RI basés sur des ontologies sont parfois confrontés à la couverture limitée de ces dernières, nous proposons de combiner la recherche sémantique et la recherche par mots clés. La suite de ce chapitre est organisée comme suit. Nous présentons tout d'abord l'architecture générale d'un modèle de RI sémantique (section 2). Ensuite, nos différentes propositions sont détaillées dans les sections qui suivent. Ainsi, deux méthodes d'extraction de concepts sont décrites dans la section 3. Ensuite, la désambiguïsation et la pondération des concepts sont détaillées en section 4. Une évaluation de ces différentes propositions est faite dans la section 5. La section 6 présente la fonctionnalité de RI associée au portail SemBiP en particulier l'approche d'expansion de requêtes et de recherche mixte concepts et mots-clés que nous avons utilisée. Enfin, la conclusion apporte quelques éléments de discussion concernant nos choix méthodologiques et les résultats obtenus.

2 Architecture d'un modèle de recherche d'information basée sur une ontologie

Dans cette section, nous présentons brièvement l'approche de RI proposée qui repose sur une ontologie de domaine. L'architecture générale d'un modèle classique de ce type est illustrée par la figure 17. Notre travail consiste à proposer des algorithmes pour la mise en œuvre des différents composants de cette architecture. L'idée est d'adapter et d'améliorer les modèles de RI classique, tels que présentés dans le chapitre 2, en exploitant les connaissances contenues dans une ressource sémantique (ontologie ou RTO). A partir de l'architecture classique de la figure 17 et afin de répondre aux différentes fonctionnalités d'un portail sémantique, nous nous sommes focalisés sur la représentation conceptuelle (indexation conceptuelle) des documents de la collection et avons réutilisé la technique classique du modèle vectoriel pour l'appariement document-requête. Nous avons également implémenté une stratégie pour l'expansion des requêtes.

2.1 L'indexation conceptuelle des documents

Le principe de l'indexation conceptuelle est de décrire les documents et les requêtes par des concepts d'une ontologie ou d'une RTO plutôt que par des mots clés comme dans la RI classique. Notre approche comprend une méthode d'identification des concepts dans les textes (section 3) et une technique de désambiguïsation (section 4). Pour la pondération des concepts identifiés, nous adaptons le modèle vectoriel (section 4). Une évaluation de notre méthode d'indexation est ensuite présentée (section 5).

2.2 La phase de recherche d'information

Dans cette étape, l'utilisateur exprime son besoin en information via une requête. Cette dernière est traitée et, à l'instar des documents, représentée par un ensemble de concepts. Sa pertinence par rapport aux documents de la collection est estimée par une fonction de correspondance. Dans ce travail, nous utilisons la mesure de cosinus qui est bien adaptée au modèle vectoriel. Ensuite, les documents retrouvés sont présentés à l'utilisateur par ordre décroissant de leur pertinence. Cette étape est illustrée avec le portail SemBiP en section 6.

3 Extraction de concepts médicaux

Notre travail s'inscrit dans les approches d'identification de concepts à base de dictionnaires (cf section 6.2.1 du chapitre 2) et propose d'explorer deux méthodes d'identification de concepts dans des textes médicaux : une méthode reposant sur un *chunker* (extracteur) pour extraire les syntagmes nominaux dans des textes avant d'identifier les concepts correspondants dans une ontologie ou RTO (3.1) et une méthode considérant des n-grammes comme termes candidats dénotant les concepts (3.2).

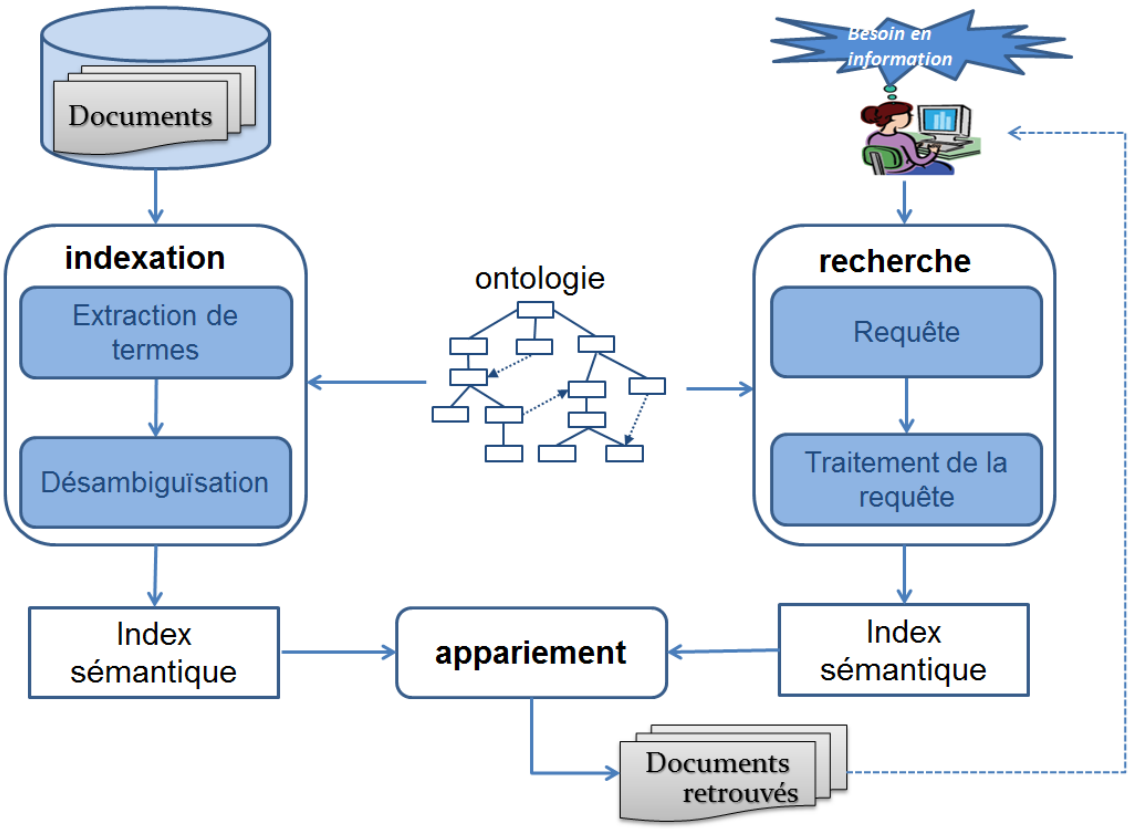


Figure 17 : Architecture classique d'un SRI sémantique

3.1 Méthode d'extraction de concepts basée sur le *chunking*

Dans cette section, nous proposons une méthode qui procède à une analyse morphosyntaxique des textes pour extraire les syntagmes nominaux. Ces derniers sont ensuite alignés aux entrées dénotant les concepts de la ressource sémantique.

Comme l'ont montré diverses expérimentations rapportées dans la littérature, la prise en compte des informations morphosyntaxiques est généralement pertinente pour la reconnaissance des entités médicales. Par exemple, MetaMap (Aronson, 2001) présentée dans le chapitre précédent et qui est une référence dans le domaine biomédical, s'appuie sur une analyse syntaxique des textes pour en extraire les concepts de l'UMLS. Dans (Ruch, 2006), l'utilisation des syntagmes nominaux pour recalculer les scores de pertinence des concepts préalablement identifiés par un système de classification de documents biomédicaux a permis également d'en améliorer significativement les performances. Dans la littérature, une large gamme d'outils dédiés à l'identification des syntagmes dans des corpus, appelés *chunkers*, ont été développés. Dans (Kang et al., 2011), les auteurs ont évalué différents chunkers sur des textes biomédicaux (corpus Genia) : GATE chunker[58], Genia Tagger[59] (Tsuruoka et al., 2005), Lingpipe[60], MetaMap (Aronson, 2001), OpenNLP[61] et Yamcha[62]. D'après leur étude,

[58] https://gate.ac.uk/
[59] http://www.nactem.ac.uk/GENIA/tagger/
[60] http://alias-i.com/lingpipe/index.html
[61] https://opennlp.apache.org/

OpenNLP donne les meilleures performances dans l'extraction des syntagmes nominaux (F-mesure de 89,7%) et verbaux (F-mesure de 95,7%) suivi de Genia Tagger et de Yamcha. Ils notent également qu'en termes de convivialité, Lingpipe et OpenNLP sont plus simples d'utilisation. Par ailleurs, cette comparaison montre que les chunkers associés aux outils spécialisés tels que MetaMap et Genia Tagger ne sont pas toujours les plus performants. Abacha et Zweigenbaum (2011) ont proposé une étude comparative de trois chunkers : MetaMap, TreeTagger-chunker[63] (Schmid, 1994) et OpenNLP. Selon cette étude, sur une comparaison basée seulement sur le rappel, TreeTagger-chunker obtient les meilleurs résultats, suivi d'OpenNLP. Cette étude montre également que MetaMap, présente quelques limites qu'un prétraitement permet de surmonter.

Afin d'élaborer notre méthode, en nous basant sur ces deux précédentes études, nous portons notre choix sur le chunker OpenNLP et l'étendons avec des traitements supplémentaires pour améliorer ses résultats. OpenNLP est une suite de modules basée sur des techniques d'apprentissage automatique pour traiter différentes tâches de TAL : segmentation des textes en phrases et en tokens, étiquetage morphosyntaxique, chunking, etc. Son module de chunking, basé sur un modèle de maximum d'entropie, permet d'identifier notamment les syntagmes nominaux, verbaux et adjectivaux dans des textes prétraités. Le maximum d'entropie est un classifieur probabiliste log-linéaire. Il se base sur la notion d'incertitude. Son principe consiste à choisir, pour un problème donné, les conclusions qui maximisent l'entropie (l'incertitude) tout en restant consistant (Jaynes, 1957). Nous proposons ainsi une méthode de repérage de concepts qui comprend deux étapes. D'abord, les syntagmes nominaux sont extraits en utilisant OpenNLP. Après traitement, ces syntagmes sont alignés aux concepts de la ressource sémantique.

3.1.1 Extraction des syntagmes nominaux

Pour annoter un texte donné, il est d'abord segmenté en phrases en utilisant le module dédié d'OpenNLP, nommément *OpenNLP Sentence Detector*. Ce module est basé sur des modèles d'apprentissage (Reynar et Ratnaparkhi, 1997) qui peuvent être entrainés sur n'importe quel corpus annoté. Il permet de vérifier si une ponctuation marque la fin d'une phrase. En effet, bien que les caractères de ponctuation marquent la fin d'une phrase, tous ne désignent pas la fin d'une phrase. Par exemple, le caractère point (« . ») peut être utilisé dans des acronymes ou encore des abréviations. Ainsi, ce sont des modèles d'apprentissage qui sont utilisés pour déterminer les fins de phrases. Le texte résultant est ensuite segmenté en tokens et les tokens sont étiquetés avec leurs catégories grammaticales correspondantes. Pour la tokenization et l'étiquetage, à l'instar de la segmentation des textes en phrases, des modèles basés sur le maximum d'entropie ont été utilisés. Enfin, le module de chunking permet, à partir de ces informations, d'identifier les syntagmes nominaux. Le tableau 9 montre un exemple de texte traité avec OpenNLP.

[62] http://www.chasen.org/~taku/software/yamcha/
[63] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

3.1.2 Alignement des syntagmes aux entrées de l'ontologie

Une fois les syntagmes déterminés, ils sont alignés aux entrées de la RTO ; on considère que la ressource utilisée couvre suffisamment les variantes et synonymes associés aux concepts. Pour cela, puisque le chunker fournit une variété de syntagmes nominaux, une phase de prétraitement est réalisée pour les filtrer. Ainsi, pour qu'un terme candidat soit retenu, il doit contenir au moins deux caractères. Les candidats constitués seulement de chiffres sont également élagués. Ensuite, pour optimiser leur correspondance, ces derniers ainsi que les entrées du vocabulaire sont d'abord normalisés. Ainsi, pour chaque terme (syntagme extrait ou entrée du vocabulaire) ayant une longueur minimale de quatre caractères, les mots vides[64] (mots non significatifs, tels que les articles et les prépositions) contenus dans ce dernier sont supprimés et les mots restants sont réduits en leur forme canonique (l'ordre des mots est conservé) grâce à une technique de lemmatisation (un algorithme de *stemming* peut être une alternative pour cette normalisation). Dans les évaluations décrites en section 5, nous utilisons la lemmatisation[65] qui s'est montrée plus efficace que le stemming sur nos tests. Les termes courts, tels que les acronymes, ne sont pas normalisés pour éviter le bruit qu'ils pourraient générer.

Tableau 9 : Exemple de texte traité avec le chunker d'OpenNLP

Texte	Motor dysfunction in mild cognitive impairment and the risk of Alzheimer disease
Syntagmes nominaux extraits	Motor dysfunction mild cognitive impairment the risk Alzheimer disease

Kang et ses collègues ont montré à travers leurs expérimentations que la plupart des erreurs de chunking sont générées par l'intégration ou l'élagage des conjonctions de coordination (« and », « or ») dans des syntagmes nominaux (Kang et al., 2011). Nous avons ainsi décomposé les syntagmes contenant ces conjonctions (« and » et « or ») en un ensemble de termes candidats. Par exemple, OpenNLP extrait le syntagme *MRSA and Serratia* et après décomposition, le terme *MRSA* est aligné au concept *MRSA - Methicillin resistant Staphylococcus aureus infection* (*C0343401*). Pour le syntagme *nausea and emesis* extrait avec cet outil, on retrouve les concepts *nausea* (C0027497) et *emesis* (C0042963). Pour optimiser les performances, nous avons également exploré une stratégie consistant à vérifier d'abord si un syntagme nominal peut être aligné tel quel à une entrée de l'ontologie avant de le décomposer. Toutefois, la décomposition de ce type de termes donne de meilleurs résultats.

Dans la section qui suit, nous présentons la seconde méthode qui consiste à utiliser les n-grammes comme unité de découpage du texte à traiter.

[64] Nous utilisons la liste disponible sur : http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T43/
[65] Nous utilisons la méthode implémentée et disponible sur http://dragon.ischool.drexel.edu/api/dragon/nlp/tool/lemmatiser/EngLemmatiser.html

3.2 Méthode d'extraction de concepts basée sur les n-grammes

La plupart du temps, l'utilisation des syntagmes nominaux et adjectivaux facilite l'identification des concepts de la RTO qui sont mentionnés dans des corpus. Toutefois, le domaine biomédical disposant d'une grande variabilité terminologique, il n'est pas simple de définir les structures morphosyntaxiques permettant de couvrir l'ensemble des termes médicaux. Par exemple, les entrées du thésaurus MeSH peuvent inclure des conjonctions comme *and*, *or*, *of*, *with*. Dans le challenge i2b2 2010 (Uzuner et al., 2011), les termes dénotant les concepts peuvent contenir des prépositions (*pain in chest, changes in mental status, tumor of the skin*), des conjonctions (*metastases in the liver and pancreas*), etc. Ceci complique la tâche d'extraction de concepts dans des textes biomédicaux. Pour prendre en compte cette variabilité, d'autres travaux ont proposé des méthodes de recherche approximative pour aligner les textes aux entrées de dictionnaires (Zhou et al., 2006a).

Dans le but de mieux couvrir ces termes complexes, nous proposons une approche basée sur les n-grammes. Par ailleurs, nous ciblons les concepts les plus spécifiques, c'est-à-dire ceux représentés par les termes les plus longs afin de réduire le bruit que peuvent entrainer des alignements partiels ou des termes imbriqués (Ruch, 2006). Par exemple, dans l'expression *Démence à corps de Lewy*, on extrait uniquement le terme *Démence à corps de Lewy*. Les termes *Corps de Lewy* et *Démence*, bien qu'apparaissant dans le texte et dénotant des concepts du vocabulaire, sont ignorés car ils sont imbriqués dans un autre terme plus spécifique. Notre hypothèse est que les termes longs sont plus spécifiques et que donc ils véhiculent des informations plus précises comparativement aux termes courts qui sont souvent plus généraux, voire vagues. En plus, les termes généraux peuvent être déduits en exploitant les relations de subsomption.

L'algorithme d'extraction de concepts que nous proposons est décrit dans l'***Algorithme 1***. Le principe est de considérer chaque n-gramme ne commençant pas ni ne terminant par un mot vide comme un candidat terme, en privilégiant les candidats les plus longs. Dans cette méthode, le texte est d'abord décomposé en phrases et chaque phrase est ensuite segmentée en tokens. L'idée est ensuite d'identifier les candidats termes maximaux correspondant à des entrées de l'ontologie. Pour cela, la longueur maximale (nombre de tokens) que peut avoir un candidat terme est fixée arbitrairement. Ensuite, à partir du début de la phrase, on extrait le candidat maximal afin de le rechercher dans la RTO. S'il est retrouvé, on passe au candidat terme suivant disjoint (lignes 25 et 26). Sinon, le sous-candidat terme gauche le plus long est considéré et recherché jusqu'à retrouver une entrée de la RTO ou atteindre le premier mot du candidat terme (ligne 12). Si un candidat terme est trouvé, on passe au candidat terme suivant. Par contre, si aucun sous-candidat n'est identifié, on avance les positions de début et fin du n-gramme (lignes 29 et 30) et la recherche redémarre avec le nouveau candidat. Ce processus est répété jusqu'à parcourir tous les tokens de la phrase.

Notons que pour la recherche des candidats termes associés à des concepts, ces derniers ainsi que les entrées de la RTO sont d'abord normalisés avant de les aligner. Ainsi, pour chaque terme, comme dans la première méthode, les mots vides contenus dans ce dernier sont

supprimés et les mots restants sont réduits chacun à leur forme canonique en utilisant une technique de lemmatisation. La longueur maximale des termes qu'on reconnait est fixée de manière arbitraire.

Algorithme 1 : TermMatcher : Extraction de termes associés aux concepts (selon les n-grammes)
Entrée: O: Ontologie, D: document, longueurMax : la longueur maximale d'un terme ;
Sortie: T: liste de termes du document D dénotant des concepts.

```
1:  Début
2:      phrases[] ← décomposer(D)
3:          pour chaque phrase dans phrases faire
4:              tokens[] ← segmenter(phrase)
5:              début ← 0
6:              taille ← taille de tokens
7:              fin ← min(début + longueurMax – 1, taille – 1)
8:              Tant que fin < taille faire
9:                  Si !estMotVide(tokens[début]) alors
10:                     dernier ← fin
11:                     trouvé ← faux
12:                     Tant que dernier ≥ début et trouvé = faux faire
13:                         Si ! estMotVide(tokens[dernier]) alors
14:                             ngram ← formerNgram(début, dernier)
15:                             normaliser(ngram)
16:                             bool ← rechercher(ngram, O)
17:                             Si bool = vrai alors
18:                                 trouvé ← vrai
19:                                 ajouter(ngram, T)
20:                             fin si
21:                         fin si
22:                         dernier ← dernier – 1
23:                     fin tant que
24:                     Si trouvé = vrai alors
25:                         début ← dernier + 1
26:                         fin ← min(début + longueurMax – 1, taille – 1)
27:                     fin si
28:                 fin si
29:                 début ← début + 1
30:                 fin ← fin + 1
31:             fin tant que
32:             Tant que début != fin faire
33:                 Si !estMotVide(tokens[début]) alors
34:                     trouvé ← faux
35:                     dernier ← fin – 1
36:                     Tant que dernier  début et trouvé = faux faire
37:                         Si !estMotVide(tokens[dernier]) alors
38:                             ngram ← formerNgram(début,dernier)
39:                             normaliser (ngram)
40:                             bool ← rechercher(ngram, O)
41:                             Si bool = vrai alors
```

```
42:                                                    trouvé ← vrai
43:                                                    ajouter(ngram, T)
44:                                                fin si
45:                                            fin si
46:                                            dernier ← dernier – 1
47:                                        fin tan que
48:                                        Si trouvé = vrai alors
49:                                            début ← dernier + 1
50:                                        fin si
51:                                    fin si
52:                                    début ← début + 1
53:                            fin tant que
54:                    fin pour
55:    fin
```

Les deux approches ci-dessus permettent, pour chaque document de la collection, d'identifier l'ensemble des termes et des concepts décrivant son contenu. Toutefois, les termes identifiés dans des textes pouvant être ambigus (i.e. associés à plusieurs concepts), nous complétons notre méthode d'extraction de concepts avec une méthode de désambiguïsation.

4 Désambiguïsation des termes et pondération des concepts

4.1 Désambiguïsation des termes

Cette étape permet, pour chaque terme ambigu, de retrouver le concept adéquat correspondant à son contexte d'utilisation. Notre méthode d'extraction de termes, en privilégiant les termes les plus longs, permet dans un premier temps de réduire les ambiguïtés mais des termes polysémiques peuvent malgré tout être retrouvés. Ainsi, nous utilisons une technique de désambiguïsation basée sur la similarité sémantique entre les concepts pour traiter les éventuels cas d'ambiguïté. L'approche de désambiguïsation que nous proposons est inspirée des travaux de (McInnes et Pedersen, 2013) et nous avons intégré un poids permettant de mesurer l'importance (pour la désambiguïsation) de chaque terme du contexte du terme ambigu. Le principe de cette méthode est, pour chaque terme ambigu, de récupérer tous les concepts qu'il dénote et de trouver le concept le plus proche sémantiquement des autres concepts de son contexte. Ce contexte est défini comme une fenêtre de texte fixée arbitrairement et peut être une phrase, un paragraphe ou même le document entier. Ainsi, pour chaque concept associé au terme ambigu, son score de similarité est estimé comme la somme du degré de similarité de ce concept avec chacun des autres concepts de son contexte. Pour tout autre terme ambigu dans le contexte, seul son concept associé le plus proche sémantiquement du concept cible (i.e., ayant le degré de similarité le plus élevé avec ce dernier) est considéré au lieu de tous les concepts qu'il dénote. A l'issue de ce calcul, le concept ayant le score maximal est retenu et assigné au terme.

Contrairement à la formule originale donnée dans (McInnes et Pedersen, 2013), nous ne considérons pas les distances (le nombre de mots les séparant) entre un terme ambigu et les

termes apparaissant dans son contexte pour le calcul des scores de similarité. De plus, la similarité entre un concept dénoté par un terme à désambiguïser et les autres concepts de son contexte est pondérée par un coefficient indiquant l'importance de leurs termes associés dans la désambiguïsation. Ce poids, pour un terme donné, est défini comme l'inverse du nombre de concepts qui lui sont associés. Nous considérons que moins un terme du contexte est polysémique, plus il est important pour la désambiguïsation des termes ambigus. Ainsi, pour les termes non ambigus, ce poids est maximal ($w(t_i) = 1$).

Formellement, soit F le contexte d'un terme ambigu t_i associé à n concepts $C_i = \{c_{i1}, \dots, c_{in}\}$ et T l'ensemble des termes de F extraits en utilisant l'***Algorithme 1*** (ou la première méthode), le score de chaque concept $c_{ij}, j \in \{1, \dots, n\}$ est défini par :

$$Score\left(c_{ij}\right) = \sum_{\substack{k=1 \\ k \neq i}}^{|T|} \max_{l \in [1,\dots,|C_k|]} (Sim_{Lin}(c_{ij}, c_{kl})) \times w(t_i)$$

avec $Sim_{Lin}(c_{ij}, c_{kl})$ la similarité sémantique entre les concepts c_{ij} et c_{kl} définie dans (Lin, 1998), et $w(t_i)$ un poids permettant d'indiquer l'importance des concepts associés au terme t_i dans la désambiguïsation. Ce poids est calculé comme suit :

$$w(t_i) = \frac{1}{nbConcept(t_i)}$$

avec $nbConcept(t_i)$, le nombre de concepts associés au terme t_i.

De cette étape, résulte un ensemble de concepts (non ambigus) qui peuvent être utilisés pour indexer le document traité. Toutefois, les concepts retrouvés dans un document ne sont pas d'importance égale pour le représenter. Dans la section qui suit, nous présentons notre schéma de pondération permettant d'associer, à chaque concept extrait, un poids indiquant sa pertinence pour le document.

4.2 La pondération des concepts

L'identification des concepts et leur désambiguïsation visent, pour chaque document d'une collection à traiter, et étant donnée une RTO, à retrouver l'ensemble des concepts qui représentent son contenu. Cependant, repérer ces concepts ne suffit pas quand on souhaite, dans le cadre de la RI, pouvoir trier les résultats à fournir selon un critère de pertinence qui dénote l'importance des concepts indexant les différents documents de la collection. C'est pourquoi des schémas de pondération sont utilisés pour déterminer le poids (l'importance) de chaque concept pour décrire le contenu d'un document. Comme nous l'avons présenté dans la section 3 du chapitre II, le modèle vectoriel (Salton et McGill, 1986) et le BM25 (Robertson et Walker, 1994) sont deux modèles de RI courants et souvent utilisés dans les campagnes d'évaluation comme des références (Goeuriot et al., 2014). Dans le cadre de ce travail, nous utilisons le schéma de pondération associé au premier (le schéma TF.IDF) et adapté aux concepts (Diallo, 2006); en termes de temps d'exécution, ce modèle est plus efficace. Ainsi, le poids d'un concept c pour un document d est défini par :

$$CF.IDF_{c,d} = \sqrt[2]{CF_{c,d}} * log\frac{N}{n_i}$$

avec $CF_{c,d}$ le nombre d'occurrences du concept c dans le document d ; N le nombre total de documents dans la collection ; et n_i le nombre de documents contenant le concept c.

Dans la section suivante, nous décrivons la méthodologie suivie pour évaluer les méthodes présentées dans la section 3.

5 Evaluation des méthodes d'extraction de concepts

L'évaluation consiste à estimer la capacité d'une méthode à identifier correctement les concepts mentionnés dans des textes. Différentes techniques d'évaluation sont proposées dans la littérature. Généralement, les résultats des systèmes sont comparés à ceux fournis par des annotateurs humains sous forme de collections de test.

Pour l'évaluation des méthodes proposées, nous suivons cette approche et utilisons les mesures classiques de rappel, précision et f-mesure.

5.1 Collections de test

Pour évaluer nos différentes propositions, nous nous servons des collections de textes standards largement utilisées dans le domaine. Ainsi, pour notre expérimentation, deux corpus de types différents sont utilisés : 1) un corpus annoté constitué de textes cliniques (ShARe/CLEF eHealth2013) et 2) un corpus composé d'articles biomédicaux extraits à partir de la base MEDLINE (corpus de Berkeley). Nous les présentons en détail dans les sous-sections suivantes.

5.1.1 Le corpus de ShARe/CLEF eHealth2013

Cette collection de textes cliniques a été fournie pour évaluer les systèmes participant à la tâche 1 du challenge ShARe/CLEF eHealth2013 (Pradhan et al., 2013; Suominen et al., 2013). L'objectif de cette tâche était d'identifier, dans un texte, les concepts médicaux appartenant au groupe sémantique ***Disorders*** de l'UMLS (cf section 4.2.1 du chapitre 3). Pour cela, les organisateurs ont fourni des corpus d'entrainement (200 documents) et de test (100 documents), composé chacun d'un ensemble de résumés de sortie de patients et de rapports d'électrocardiogramme, d'échocardiographie et/ou de radiologie en texte libre.

Chaque document dans la collection d'entrainement est annoté et l'ensemble des entités qui dénotent des concepts de type « disorder » sont identifiées. L'annotation a été réalisée par deux annotateurs professionnels formés pour cette tâche, suivie d'une étape d'ajustement. Un concept est considéré comme un « disorder » s'il est contenu dans la terminologie SNOMED-CT et appartient à un des types sémantiques suivants de l'UMLS (groupe sémantique ***Disorders***) : Congenital Abnormality, Acquired Abnormality, Injury or Poisoning, Pathologic Function, Disease or Syndrome, Mental or Behavioral Dysfunction, Cell or Molecular Dysfunction, Experimental Model of Disease, Anatomical Abnormality, Neoplastic Process et

Sign or Symptom. La figure 18 présente un exemple de document annoté. Dans le cadre de notre évaluation, l'objectif est simplement de mesurer la capacité de nos méthodes à identifier correctement les concepts de SNOMED-CT mentionnés dans les textes fournis. Puisque les méthodes proposées sont non supervisées, nous utilisons l'ensemble d'entrainement pour les évaluer.

00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0037199||337||346
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||398||411
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||557||569
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0037199||651||660
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||873||885
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0037199||896||905
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||1059||1069
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0263978||1143||1165
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0426460||1226||1238||1280||1284
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||C0549397||1226||1247
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||1363||1374
00587-400001-RADIOLOGY_REPORT.txt||Disease_Disorder||CUI-less||1531||1555

Figure 18 : Exemple d'annotation dans le corpus de ShARe/CLEF eHealth2013. La première colonne désigne le document, la deuxième le type d'annotation, la troisième le CUI du concept s'il est contenu dans SNOMED-CT, CUI-less sinon. Les dernières colonnes marquent les positions de début et fin des entités dans le document

Les concepts de SNOMED-CT appartenant au groupe sémantique ***Disorders***, utilisés pour l'évaluation, sont extraits à partir de l'UMLS (version 2012AA). Ils représentent 88 092 concepts correspondant à 528 233 termes distincts de l'UMLS. Ces derniers forment donc notre dictionnaire d'entrées.

5.1.2 Le corpus de Berkeley

Créé dans le but d'extraire les relations sémantiques entre les concepts médicaux de types « maladie » et « traitement », le corpus de Berkeley constitué de titres et résumés d'articles scientifiques extraits à partir de la base MEDLINE a été annoté par un étudiant en Master ayant des connaissances en biologie (Rosario et Hearst, 2004). Ce dernier a parcouru le corpus, phrase par phrase, et identifié les différents types de relations existant entre les concepts de types « maladie » et « traitement ». La figure 19 montre un exemple d'annotation qui indique que le médicament *dexfenfluramine hydrochloride* est un « traitement » de la « maladie » *obesity*. Notons qu'aucune convention d'annotation n'est spécifiée. Par exemple, dans l'expression *ovarian cancer*, pour certaines phrases, seul *cancer* est identifié comme une « maladie » tandis que dans d'autres, le groupe de mots *ovarian cancer* est désigné comme une « maladie ». Par ailleurs, l'annotation est faite indépendamment des structures syntaxiques des entités ; les déterminants peuvent ainsi être inclus ou non dans les entités.

Dans notre évaluation, nous nous intéressons à l'identification de ce type de concepts dans le corpus de Berkeley. Ce dernier étant composé de 3 654 phrases annotées par 3 364 entités médicales dont 2 468 distinctes. Un alignement exact avec les entrées de l'UMLS a permis de retrouver uniquement 1 218 (49,4%) de ces entités dans la ressource. En plus, ces dernières sont réparties dans 42 types sémantiques qui englobent plus d'un million de concepts. Dans notre évaluation, nous considérons l'ensemble des termes utilisés pour annoter les documents comme notre vocabulaire d'entrée. Les entités médicales cibles sont, par conséquent, restreinte à cet ensemble. Notons que c'est une restriction très forte mais permet globalement d'évaluer nos méthodes d'identification de concepts.

<DIS> Obesity </DIS> is an important clinical problem and the use of <TREAT> dexfenfluramine hydrochloride </TREAT> for weight reduction has been widely publicized since its approval by the Food and Drug Administration.

Figure 19. Exemple d'annotation dans le corpus de Berkeley

5.2 Les métriques d'évaluation

Pour l'évaluation, nous nous servons des mesures classiques utilisées habituellement en RI : rappel, précision et f-mesure que nous définissons ci-dessous suivant notre contexte d'utilisation. Soit RET, le nombre d'annotations retournées par le système, COR, le nombre d'annotations correctes parmi celles retournées et REF, le nombre d'annotations de référence, les mesures de rappel, précision et f-mesure sont définies respectivement comme suit :

$$R = \frac{COR}{REF}$$

$$P = \frac{COR}{RET}$$

$$F = \frac{2 \times P \times R}{P + R}$$

Ici, nous avons appliqué une comparaison exacte entre les concepts identifiés par le système et ceux de l'ensemble de référence. Nous avons implémenté ces mesures qui sont parmi celles utilisées dans la campagne ShARe/CLEFeHealth 2013.

Vu ses bonnes performances dans (Kang et al., 2011) et sa spécificité au domaine biomédical, les résultats de Genia Tagger (qui permet aussi un d'identifier des entités médicales dans des textes anglais) sont présentés et comparés à ceux fournis par nos méthodes.

Dans l'expérimentation sur le corpus de ShARe/CLEF eHealth2013, nous ne considérons pas les annotations issues d'entités disjointes (i.e., celles composées de tokens non contigus) car ces dernières ne sont pas prises en compte dans les différentes méthodes que nous évaluons et comparons. Par ailleurs, les performances des différentes méthodes sont estimées sans et avec l'utilisation d'une méthode de désambiguïsation. Dans le premier cas, tous les concepts associés à un terme sont considérés dans l'évaluation du système tandis que dans le deuxième,

chaque terme identifié dénote un seul concept. Pour le calcul de la similarité sémantique entre les concepts dans la phase de désambiguïsation, nous avons utilisé le package Perl *UMLS::Similarity* (McInnes et al., 2009).

5.3 Résultats

Dans cette partie, nous utilisons les dénominations suivantes pour les différents systèmes que nous avons comparés (les deux derniers étant ceux implémentant nos propositions) :

- **Genia** : le système basé uniquement sur Genia Tagger ;
- **OpenNLP** : le système basé sur le chunker d'OpenNLP ;
- **Genia+** : le système Genia avec une décomposition des syntagmes nominaux contenant des conjonctions ;
- **OpenNLP+** : le système implémentant notre première méthode d'extraction de concepts ;
- **Ngram+** : le système implémentant notre deuxième méthode basée sur les n-grammes.

5.3.1 Résultats sur le corpus de ShARe/CLEF eHealth2013

Dans un premier temps, l'identification des concepts se fait sans la désambiguïsation. Les résultats des différentes méthodes sur les 200 documents de l'ensemble d'entrainement de ShARe/CLEF eHealth2013 sont présentés dans le tableau 10.

Tableau 10 : Résultats des différents systèmes sans désambiguïsation sur le corpus ShARe/CLEF eHealth2013

Système	Précision	Rappel	F-mesure
Genia	0,54	0,36	0,43
OpenNLP	0,57	0,34	0,42
Genia+	0,55	0,41	0,47
OpenNLP+	**0,58**	0,38	0,46
Ngram+	0,43	**0,76**	**0,55**

La méthode basée sur les n-grammes obtient les meilleurs résultats avec une f-mesure de 0,55. Cette dernière permet d'identifier une bonne partie des concepts cibles dans les textes (rappel de 0,76) mais avec une précision modeste (0,43). Concernant les systèmes utilisant le chunking, OpenNLP, qui avait pourtant obtenu les meilleurs résultats dans des expérimentations antérieures (Kang et al., 2011; Abacha et Zweigenbaum, 2011), est légèrement dépassé par Genia Tagger. La décomposition des syntagmes nominaux contenant des conjonctions (« and » et « or ») permet également d'améliorer les résultats : une f-mesure de 0,42 à 0,46 et de 0,43 à 0,47 pour respectivement OpenNLP et Genia Tagger.

Dans un deuxième temps, une phase de désambiguïsation où un terme est associé à un concept unique est intégrée dans le processus d'identification des concepts (Tableau 11). Pour OpenNLP+, la désambiguïsation, bien qu'augmentant sa précision, ne permet pas d'améliorer les performances globales à cause d'une diminution du rappel. En revanche, pour Ngram+, nous notons une nette amélioration des résultats avec une augmentation de la précision (de

0,43 à 0,48) mais son rappel diminue légèrement. Les systèmes partent d'un déséquilibre rappel/précision différent; si la précision est inférieure au rappel, la désambigüisation permet d'améliorer les résultats en augmentant la précision; sinon elle ne permet pas d'augmenter les performances (f-mesure).

5.3.2 Résultats sur le corpus de Berkeley

L'évaluation sur le corpus de Berkeley révèle de meilleures performances. Le tableau 12 montre les résultats des différents systèmes. Ngram+ obtient les meilleures performances avec une f-mesure de 0,86. Elle est suivie de OpenNLP+. Dans cette expérimentation, la décomposition des syntagmes nominaux contenant des conjonctions *and* ou *or* permet, comme dans la première, d'améliorer les performances globales des systèmes même si leurs précisions baissent. Nous constatons également qu'OpenNLP donne de meilleurs résultats que Genia Tagger dans ce corpus qui est pourtant un outil spécialisé du domaine.

Tableau 11 : Résultats des différents systèmes avec désambiguïsation sur le corpus ShARe/CLEF eHealth2013

Système	Précision	Rappel	F-mesure
Genia	0,60	0,33	0,43
OpenNLP	0,65	0,31	0,42
Genia+	0,62	0,38	0,47
OpenNLP+	**0,66**	0,35	0,46
Ngram+	0,49	**0,72**	**0,58**

Tableau 12 : Résultats des différents systèmes sur le corpus de Berkeley

Système	Précision	Rappel	F-mesure
Genia	0,79	0,65	0,72
OpenNLP	0,83	0,65	0,73
Genia+	0,78	0,68	0,72
OpenNLP+	**0,82**	0,67	0,74
Ngram+	0,78	**0,96**	**0,86**

5.4 Analyse des résultats

Les expérimentations réalisées sur deux corpus différents ont permis d'obtenir des résultats intéressants. Deux techniques non supervisées sont explorées pour l'extraction des concepts à partir de documents biomédicaux : une technique reposant sur le chunking et une autre basée sur les n-grammes.

Une évaluation sur le corpus clinique montre que la première approche permet d'obtenir des résultats plus précis (précision = 0,57) mais reste limitée en termes de rappel (0,38). La deuxième méthode, considérant tout n-gramme comme un terme candidat pouvant dénoter un concept, permet de retrouver une bonne partie des concepts mentionnés dans le texte (rappel =

0,76) mais avec moins de précision (précision = 0,43). Globalement, les performances de cette dernière restent meilleures avec une f-mesure de 0,55 contre 0,46 pour la première.

L'intégration d'une phase de désambiguïsation a permis d'accroître la précision des différents systèmes. On note cependant une baisse du rappel qui est due à des erreurs de désambiguïsation. Même si elle n'a pas permis d'augmenter les performances de l'approche basée sur le chunking, la désambiguïsation améliore clairement les résultats (i.e. la f-mesure) de la méthode reposant sur les n-grammes. Ceci peut s'expliquer par le fait qu'avec le chunking, on a moins de candidats et donc moins de bruit. Ainsi, une étape de désambiguïsation reste moins cruciale comparée à la méthode utilisant les n-grammes qui génère beaucoup de termes candidats ambigus.

Compte tenu de la comparaison de ces deux approches sur le corpus de ShARe/CLEF eHealth2013, nous avons noté que l'utilisation d'une technique de chunking pour l'extraction de concepts dans des textes cliniques assure plus de précision mais entraine beaucoup de faux négatifs (faible rappel) ; trop de concepts sont ignorés. Ce résultat confirme les règles suivies pour l'annotation de cette collection de test où on ignore la structure syntaxique des termes. Nous avons remarqué également la faible précision de la deuxième méthode, qui peut s'expliquer par le type de corpus utilisé pour l'évaluation. En effet, ce dernier étant composé de dossiers patients, il contient beaucoup d'acronymes et d'abréviations dont l'approche basée sur les n-grammes peine à identifier correctement les concepts associés. La figure 20 compare les résultats des deux méthodes avec et sans désambiguïsation sur le corpus ShARe/CLEF eHealth2013. Vu que nous avons ignoré les entités disjointes, une comparaison directe avec les résultats des différents systèmes participant au challenge n'est pas simple. En effet, la reconnaissance d'entités disjointes est particulièrement délicate et explique les moins bonnes performances des autres systèmes comparativement aux nôtres. Cependant, le meilleur résultat rapporté pour cette tâche est une f-mesure de 0,59 (Pradhan et al., 2013). La non-prise en compte des entités disjointes dans l'évaluation impacte aussi la précision de notre deuxième méthode. En effet, cette dernière, à la place des entités les plus spécifiques, retrouve des sous-entités imbriquées dans ces dernières, ce qui diminue sa précision. Nous notons également que Genia Tagger, qui est un outil spécialisé, dépasse légèrement notre première méthode basée sur OpenNLP.

Dans une deuxième expérimentation sur un corpus d'articles scientifiques (corpus de Berkeley), ces différentes méthodes obtiennent de meilleures performances. Nous devons toutefois préciser que les concepts à identifier sont restreints à ceux utilisés pour annoter les documents de la collection. De ce fait, la mesure de précision reste moins pertinente pour cette évaluation. Ainsi, en se basant seulement sur le rappel, notre méthode utilisant les n-grammes reste plus performante en permettant de retrouver presque (rappel = 0,96) toutes les entités médicales mentionnées. Nous rappelons, comme évoqué dans sa description, qu'aucune règle n'est suivie pour l'annotation de cette collection. Ceci peut expliquer le rappel plus faible obtenu par notre première méthode basée sur le chunking comparée à la deuxième qui est plus flexible.

Dans les deux collections, la décomposition des syntagmes nominaux contenant des conjonctions « and » ou « or » s'est montrée très pertinente et a permis d'améliorer significativement les résultats.

Dans la prochaine section, nous mettrons en œuvre ces différentes propositions pour la mise en place du portail SemBiP.

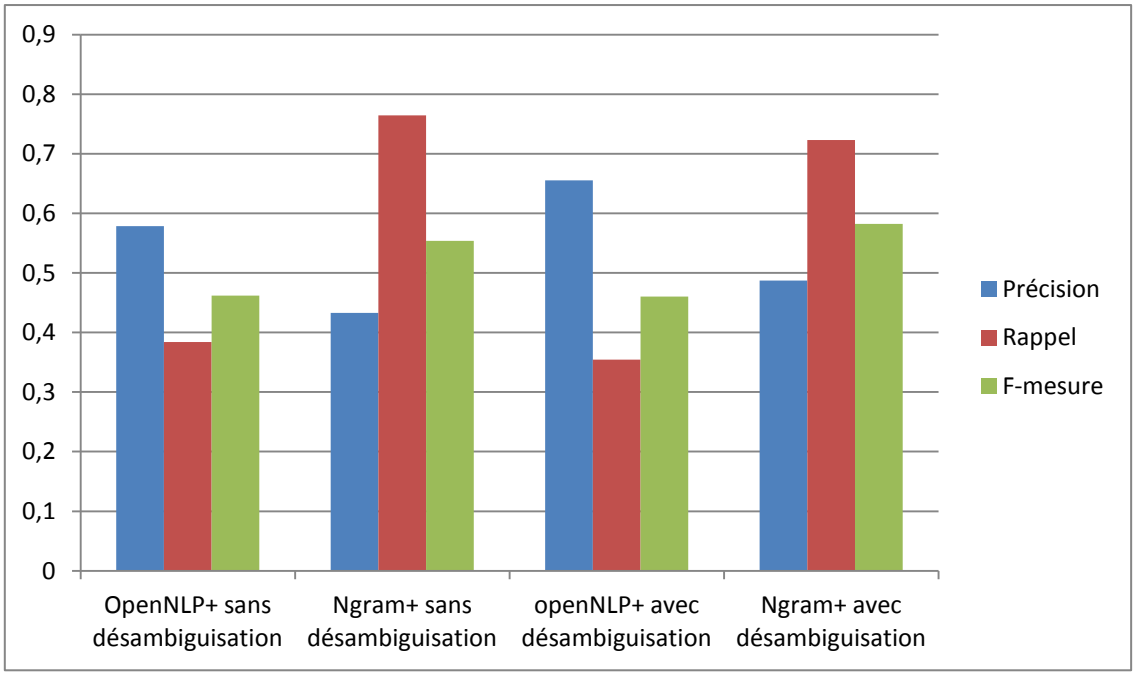


Figure 20 : Comparaison des deux méthodes OpenNLP+ et Ngram+ avec et sans désambiguïsation sur le corpus clinique de ShARe/CLEF eHealth2013

6 Application pour la mise en œuvre du portail SemBiP

L'objectif du portail SemBiP[66] est de permettre à différents utilisateurs, tels que des étudiants, médecins généralistes ou spécialistes, chercheurs et des usagers du grand public, d'accéder facilement à une synthèse critique de la littérature mondiale de référence sur la maladie d'Alzheimer et les syndromes apparentés, grâce au travail de lecture critique effectué pour alimenter la base bibliographique BiblioDem. Ce portail, que nous avons conçu et mis en œuvre, contribue à une meilleure connaissance de la maladie et à suivre son évolution pour une meilleure prise en charge des patients. Le contexte bilingue de BiblioDem (résumé en anglais et analyse critique correspondante en français) nécessite de pouvoir accéder aux ressources en ayant la possibilité d'effectuer des recherches aussi bien en français qu'en anglais. L'ontologie bilingue OntoAD, présentée au chapitre 3, est utilisée pour supporter le portail : indexation conceptuelle des documents, autocomplétion avec les termes associés aux concepts, etc.

[66] http://lesim.isped.u-bordeaux2.fr/sembip3.0/

6.1 Le portail SemBiP

Dans le portail SemBiP, interviennent différents acteurs : 1) un administrateur qui assure la gestion des différentes ressources, 2) les relecteurs qui se chargent de réaliser les analyses critiques des articles sélectionnés, et 3) les utilisateurs de différents types (médecins, étudiants, grand public, etc.) qui peuvent réaliser des recherches sur le portail. Différentes tâches doivent ainsi pouvoir se réaliser via SemBiP : l'attribution des documents aux relecteurs pour leur analyse, la validation et l'indexation des documents analysés et la fonction principale qui est l'accès à la bonne information. Dans ce qui suit, nous présentons le module de RI du portail. Notons que toute la partie « backend » du système pour l'administration a également été implémentée, notamment la récupération des articles liés à la maladie d'Alzheimer à partir de la base MEDLINE, l'affectation des articles à analyser aux experts ou encore la génération du bulletin bibliographique BiblioDémences.

Pour la mise en œuvre de SemBiP, nous avons exploité les algorithmes décrits précédemment implémentés en Java et disponibles sous forme d'une bibliothéque réutilisable. Au sein de SemBiP sont mises en œuvre :

- Une indexation conceptuelle automatique des ressources du portail basée sur l'ontologie OntoAD (en utilisant l'***Algorithme 1*** pour l'identification des concepts), complétée par une indexation en texte libre, en utilisant la bibliothéque open source Apache Lucene[67] ;
- Une aide à la saisie de requête pour aider l'utilisateur à exprimer ses besoins grâce à l'implémentation d'une technique d'auto-complétion qui utilise les termes associés aux concepts de l'ontologie OntoAD;
- Une fonctionnalité de surlignage lors de la présentation des résultats, de manière à ce que l'utilisateur comprenne directement la raison pour laquelle tel ou tel document lui sont retournés.

Nous détaillons dans la section suivante, la fonctionnalité de recherche et l'expansion sémantique telles que mises en œuvre dans le portail.

6.2 La phase de recherche d'information

Dans la phase de recherche, l'utilisateur exprime son besoin en information sous la forme d'une requête. Cette dernière peut être exprimée en texte libre ou via un formulaire où l'utilisateur peut sélectionner facilement les concepts correspondant à son besoin. Dans le premier cas, sa requête est analysée et les concepts correspondants sont identifiés. Dans le deuxième cas, l'utilisateur sélectionne lui-même les concepts qui vont constituer sa requête.

Après le traitement de la requête de l'utilisateur, cette dernière est soumise au SRI, qui retourne l'ensemble des documents jugés pertinents pour la requête. Ainsi, chaque document retrouvé est associé à un score qui représente sa pertinence par rapport à la requête. Pour cela, à l'instar des documents, la requête est d'abord représentée par un vecteur de concepts.

[67] http://lucene.apache.org

Ensuite, la mesure du cosinus est utilisée pour calculer sa pertinence pour chaque document de la collection. Les documents sont retournés par ordre décroissant de leur pertinence.

Pour améliorer les performances des SRI, différentes stratégies permettant d'optimiser la correspondance entre les documents et les requêtes peuvent être envisagées. Dans ce travail, nous proposons d'implémenter deux techniques : une technique d'expansion de requêtes exploitant la similarité sémantique (6.2.1) et une technique de combinaison de la recherche sémantique et la recherche par mots clés (6.2.2).

6.2.1 Expansion de requêtes

Dans un premier temps, nous avons implémenté une technique d'expansion de requêtes basée sur la hiérarchie des concepts d'une RTO. Pour cela , la requête est étendue par les concepts enfants (sous-concepts directs) des concepts qu'elle contient (Dramé et al., 2014). Par exemple, une requête avec le concept *déficience cognitive* peut être étendue en incluant des concepts comme *déclin cognitif lié à l'âge* et *trouble de la mémoire*, qui sont potentiellement d'intérêt puisqu'ils sont des sous-concepts de *déficience cognitive*.

Une évaluation de cette technique sur une collection de test (voir section 6.4.1) montre que tous les sous-concepts ne sont pas pertinents pour étendre une requête.

Ainsi, au lieu de tous les concepts enfants, la requête est étendue avec seulement les concepts enfants sémantiquement proches de ceux qu'elle contient. Dans ce travail, nous utilisons la mesure de similarité sémantique proposée dans (Lin, 1998). Par exemple, bien que les concepts *Maladie d'Alzheimer* et *Démence vasculaire* soient des sous-concepts directs de *Démence*, le premier est plus proche avec un degré de similarité de 0,96 contre 0,79 pour le second. En effet, la mesure de Lin utilise les contenus informationnels des concepts, permettant d'estimer leur spécificité, pour déterminer leur similarité; deux concepts fils d'un même concept mais avec des contenus informationnels différents ont des similarités différentes avec leur père. Ainsi, l'expansion d'une requête contenant le concept *Démence* par le concept *Maladie d'Alzheimer* devient plus pertinente. Pour matérialiser ceci, les nouveaux concepts sont associés à des poids en fonction de leur proximité aux concepts originaux de la requête. En nous inspirant du travail de (Hliaoutakis et al., 2006), nous calculons ce poids comme suit :

$$q_j = \frac{1}{|Q|} \sum_{c_i \in Q}^{i \neq j} sim(c_i, c_j)$$

avec Q la requête originale, $sim(c_i, c_j)$ la similarité entre les deux concepts c_i et c_j, et c_j un concept appartenant à l'ensemble formé par les hyponymes des concepts de la requête.

Bien que notre ontologie modélise une bonne partie des connaissances du domaine d'application, un test avec les utilisateurs (nous avons sollicité deux spécialistes pour tester de manière préliminaire le module de recherche) a montré que, pour exprimer leurs besoins en information, ces derniers utilisent également des termes qui ne sont pas dans l'ontologie car ils ne sont pas spécifiques du domaine d'application mais qui sont pourtant utiles. Par exemple, les termes *incidence*, *cohorte* ou encore *régime alimentaire* sont utilisés pour

rechercher des documents concernant la maladie d'Alzheimer. Pour faire face à cette situation, nous avons proposé, dans un deuxième temps, de combiner la recherche sémantique et la recherche par mots clés.

6.2.2 Combinaison de la recherche sémantique et de la recherche par mots clés

Beaucoup de travaux se sont intéressés à la combinaison de la recherche sémantique et la recherche par mots-clés (Bhagdev et al., 2008; Castells et al., 2007), qualifiée d'approche hybride. Une technique simple et couramment utilisée consiste à définir la fonction de correspondance de la méthode hybride comme une combinaison linéaire de celles des systèmes la composant. Par exemple, dans (Fernández et al., 2011), la fonction de correspondance du système hybride entre un document d et une requête q est définie par :

$$hSim(d,q) = \gamma\, semSim(d,q) + (1-\gamma)\, kSim(d,q)$$

avec $semSim\,(d,q)$ le score de pertinence entre le document d et la requête q fourni par la méthode sémantique, $kSim\,(d,q)$ celui obtenu avec la recherche par mots clés et γ une constante comprise entre 0 et 1.

Dans notre cas, nous considérons la recherche par mots clés comme une méthode complémentaire. Ainsi, si la requête est constituée seulement de concepts de l'ontologie, une recherche sémantique pure est effectuée. Dans le cas où elle ne contient aucun concept, notre système devient un moteur de RI classique basé sur les mots clés. Dans le cas où la requête est constituée de concepts et de mots clés (non vides), une nouvelle requête est générée en considérant l'ensemble des concepts et les mots comme des clauses d'une requête booléenne. Par exemple, avec la requête « incidence de la démence », on génère la nouvelle requête **Keywords** : *incidence* AND **Semantic** : *Démence* (CUI C0497327). Cette nouvelle requête booléenne permet la recherche du mot *incidence* en utilisant la recherche par mots clés (le champ **Keywords**) et la recherche du concept « *démence* » dans l'index sémantique (le champ **Semantic**) puis de combiner leurs résultats Notons que, contrairement au modèle booléen classique, les résultats sont retournés par ordre décroissant de pertinence.

Figure 21 : Exemple d'auto-complétion guidant l'utilisateur dans la formulation de ses requêtes

6.3 Implémentation de l'interface

L'interface d'accueil du portail SemBiP permet à l'utilisateur de choisir la langue dans laquelle il souhaite écrire sa requête (la langue par défaut étant le français). En fonction de la langue choisie et de la séquence de caractères qu'il commence à saisir, une liste de termes commençant par cette séquence et dénotant les concepts de l'ontologie lui est suggérée (Figure 21). Plus précisément, les concepts correspondant à ce que saisit l'utilisateur sont proposés sous forme de liste déroulante. Ainsi, à l'aide de cette technique d'auto-complétion, l'utilisateur peut formuler facilement sa requête en utilisant les concepts ayant servi à indexer les documents de BiblioDem. Comme illustré dans la figure 22, cette requête peut aussi être complétée par des mots clés (dans la langue choisie) dans le cas où les concepts seuls ne permettent pas à l'utilisateur d'exprimer son besoin en information. Il doit enfin sélectionner les champs sur lesquels il souhaite effectuer sa recherche parmi : titre, résumé, analyse ou sur tous les champs (option par défaut).

Figure 22 : Combinaison de la recherche par mots clés et de la recherche sémantique

Dans le cas où aucun mot clé n'est saisi pour compléter la requête, une recherche sémantique pure, basée sur les concepts de l'ontologie, est réalisée. Si l'utilisateur souhaite, par contre, faire une recherche basée seulement sur les mots clés, il ignore tout simplement le champ de saisie réservé aux concepts. Dans le cas où il effectue une recherche sémantique couplée à une recherche par mots clés, la méthode présentée dans la section 6.2 est utilisée pour estimer le degré de correspondance entre sa requête et les documents de la collection.

La présentation des résultats se fait dans l'ordre décroissant de pertinence des documents retrouvés. Par ailleurs, les concepts correspondant à la requête et apparaissant dans les

documents retournés (et plus précisément dans le titre, le résumé, et/ou l'analyse en fonction du champ sur lequel l'utilisateur a choisi d'effectuer sa recherche) sont surlignés (Figure 23). Ceci permet à l'utilisateur de comprendre pourquoi un tel document lui est retourné par le système.

La recherche peut également se faire en utilisant les mots clés qui ont été associés aux documents manuellement par une documentaliste. Pour cela, un autre formulaire a été développé (accessible via l'onglet ***Recherche par Mots Clés***) où les documents peuvent être recherchés en fonction des mots clés qui leur ont été assignés manuellement, de leurs auteurs et/ou de leur année de publication.

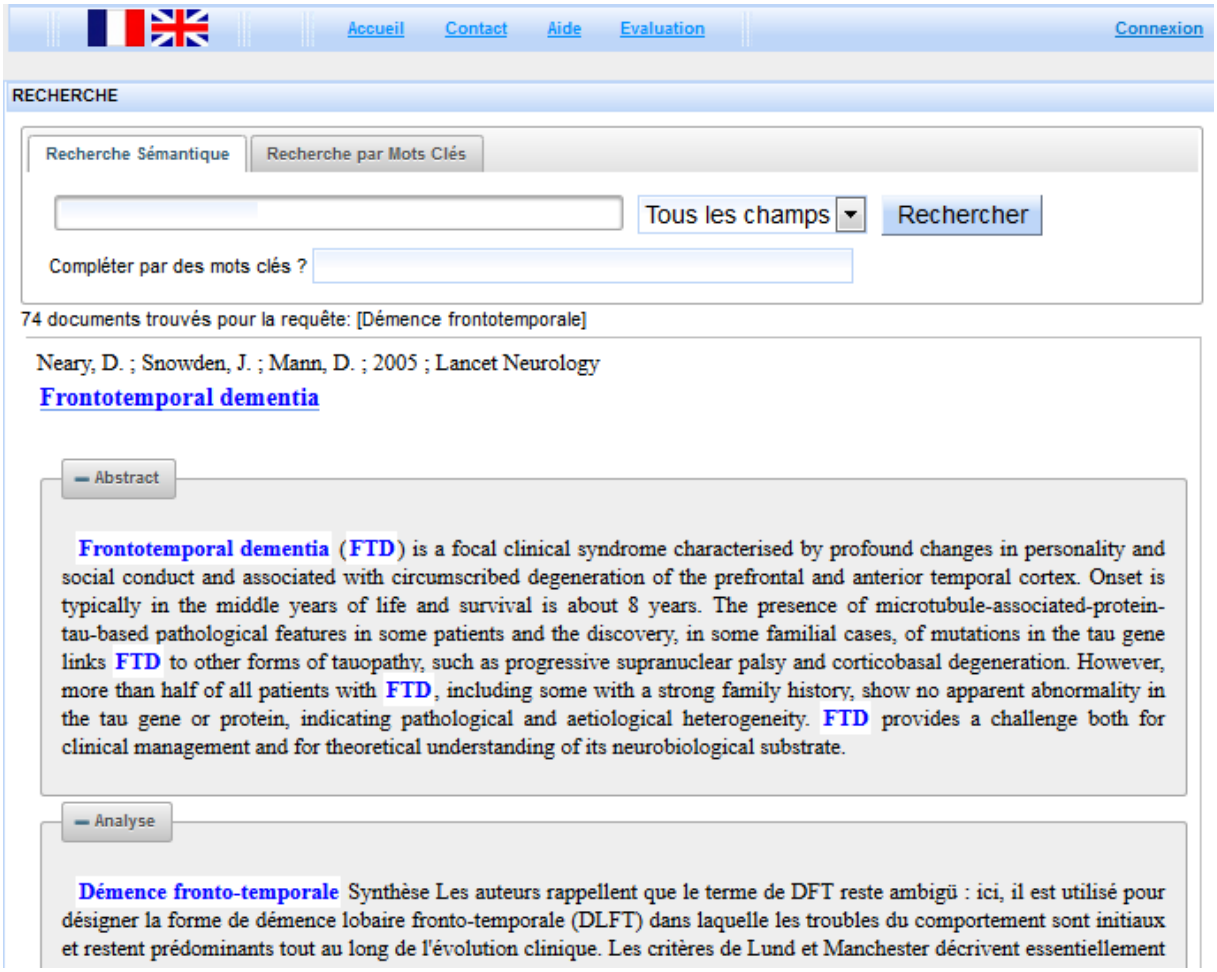


Figure 23 : Surlignage des termes dénotant les concepts de la requête

6.4 Evaluation du portail SemBiP

Pour mesurer la pertinence des services du portail SemBiP, nous avons réalisé une évaluation système pour l'efficacité de la stratégie d'expansion de requête et une évaluation orientée utilisateur.

6.4.1 Evaluation système

La technique d'expansion basée sur la hiérarchie a été évaluée sur une large collection[68] (plus d'un million) de documents fournie dans le cadre du challenge CLEF/eHealth 2014 (Goeuriot et al., 2014). Les requêtes ne pouvant être exprimées complètement par des concepts de l'UMLS, nous avons proposé de les étendre par des termes proches. Pour cela, nous avons proposé trois configurations (Runs) : 1) le Run 1, un système de RI classique basé sur le modèle vectoriel, considéré comme notre baseline, 2) le Run 2 où les termes de la requête sont étendus avec leurs synonymes extraits de l'UMLS et les termes associés à leurs sous-concepts et 3) le Run 3 où la requête est étendue seulement avec les synonymes des termes qu'elle englobe (Dramé et al., 2014). Les résultats sont présentés dans le tableau 13. Bien que l'expansion ait permis d'améliorer les performances du SRI, le Run 2 utilisant les sous-concepts a donné de moins bons résultats que le Run 3. Ceci nous a permis de montrer que l'exploitation de tous les sous-concepts n'est pas toujours appropriée et nous avons ainsi proposé une méthode d'expansion de requêtes basée sur la similarité sémantique à la place.

Tableau 13 : Résultats de nos différents runs sur la collection de la tâche 3 du CLEF eHealth 2014 ; Run 1 (TF.IDF), Run 2 (Expansion avec synonymes et sous-concepts), Run 3 (Expansion avec synonymes seulement).

Run	P@5	P@10	NDCG@5	NDCG@10
Run 1	0,50	0,51	0,50	0,50
Run 2	0,54	0,53	0,55	0,54
Run 3	0,57	0,55	0,57	0,56

6.4.2 Evaluation orientée utilisateur

Un questionnaire comprenant les questions suivantes a été soumis aux utilisateurs :

- Trouvez-vous intéressant de mettre en place un portail sémantique dédié à la maladie d'Alzheimer ?
- Le portail SemBiP est-il facile à utiliser ?
- Comment jugez-vous les résultats du système de recherche d'information ?
- Comment trouvez-vous les performances du système de recherche d'information en termes de rapidité ?
- L'auto-complétion au moment de la saisie facilite-t-elle l'expression de votre besoin en information ?
- La mise en valeur des termes de la requête dans le document permet-elle d'avoir une idée sur la pertinence des résultats ?
- Le vocabulaire utilisé pour supporter ce portail couvre-t-il suffisamment le domaine ?
- Quelles fonctionnalités souhaiteriez-vous qu'on améliore ou intègre au portail ?

[68] Fournie par les organisateurs de la tâche 3 du CLEF/eHealth2014 : http://clefehealth2014.dcu.ie/task-3

En répondant à ces questions, les utilisateurs donnent des jugements gradués de 1 (non utile ou non pertinent) à 5 (très utile ou très pertinent). L'évaluation est toujours en cours mais les premiers retours sont positifs. Sur huit utilisateurs qui ont participé à l'évaluation, cinq trouvent la mise en place du portail très important (score 5) et les trois autres intéressant (score 4). Cinq utilisateurs notent que le portail est facile à utiliser tandis que trois trouvent qu'il est un peu complexe. Ils trouvent globalement (6) que les résultats du système sont pertinents tandis que deux sondés trouvent ces résultats moyennement pertinents (score 3). En termes de rapidité, le système est aussi jugé très efficace. Pour sept des huit utilisateurs, l'auto-complétion facilite l'expression de leur requête. Ils jugent cependant la mise en valeur des termes de la requête moins utile (score 3 pour six utilisateurs). La couverture de l'ontologie est jugée globalement satisfaisante. La présentation des résultats est également jugée appropriée. Les sondés ont aussi suggéré différentes pistes pour rendre le portail plus convivial et adapté :

- Une expression de requêtes booléennes où l'utilisateur lui-même peut lier les concepts par des opérateurs booléens (OU, ET et NON) de son choix ;
- Une recherche plus flexible permettant aux utilisateurs de choisir leurs propres critères de pertinence (thématique, auteur, année, etc.) ;
- Une navigation intuitive entre les documents similaires thématiquement.

Nous avons eu globalement des retours positifs sur le portail et les utilisateurs portent un grand intérêt à cette application. Avec une évaluation élargie dans le futur, les fonctionnalités du portail SemBiP pourront être améliorées pour être un véritable outil d'information sur la maladie d'Alzheimer. Nous envisageons d'améliorer notamment la présentation des résultats. Ainsi, au lieu d'une liste des documents résultats ordonnée par pertinence, les documents pourront être présentés à l'utilisateur sous forme de groupes « cohésifs » où ceux traitant des sujets similaires sont regroupés ensemble (Renoust et al., 2013). L'utilisateur aura ainsi une visualisation plus intuitive des résultats.

7 Conclusion

Dans ce chapitre, nous avons présenté une approche de RI sémantique (RIS) guidée par une ontologie. Nous nous sommes intéressés à différentes questions soulevées par la RI sémantique : le repérage de concepts dans des corpus, la désambiguïsation de termes basée sur des mesures de similarité sémantique, la pondération des concepts, l'expansion de requêtes et l'incomplétude des ressources sémantiques.

Concernant l'extraction des concepts, nous avons exploré deux techniques différentes : une utilisant le chunking et une autre basée sur les n-grammes. D'après l'évaluation sur deux corpus de textes largement utilisés dans la communauté RI (un corpus de textes cliniques et un corpus constitué d'articles scientifiques), la deuxième approche a montré de meilleurs résultats. Nous avons aussi étudié l'impact de la désambiguïsation sur l'identification des concepts médicaux. Cette dernière, même si elle reste moins intéressante pour la méthode utilisant le chunking, a permis d'améliorer significativement les performances de la méthode

basée sur les n-grammes. Une fois les concepts identifiés, le schéma TF.IDF a été utilisé pour estimer leur poids correspondant. Ces différentes propositions ont été appliquées pour la mise en œuvre du portail sémantique SemBiP, dédié à la maladie d'Alzheimer et aux syndromes apparentés, reposant sur l'ontologie OntoAD, présentée dans le chapitre précédent. Pour améliorer l'appariement entre les documents et les requêtes, nous avons implémenté une technique d'expansion de requêtes basée sur la similarité sémantique entre les concepts. Enfin, pour faire face à la possible incomplétude de la RTO, nous avons proposé la combinaison de la recherche sémantique et la recherche par mots clés; l'évaluation de celle-ci reste une des perspectives de ce travail. Le portail SemBiP est aujourd'hui opérationnel et accessible à tout utilisateur souhaitant s'informer sur la base de faits scientifiques. Les premiers résultats de l'évaluation de ce portail sont prometteurs.

Notre méthode d'indexation conceptuelle s'appuie principalement sur le contenu textuel des documents pour identifier les concepts pertinents permettant de les représenter. Toutefois, les textes intégraux des documents ne sont pas toujours accessibles notamment dans le domaine biomédical où parfois seuls les titres et résumés sont disponibles. L'indexation des documents complets à partir de ces informations partielles reste aujourd'hui un défi. Nous nous intéressons à cette problématique dans le chapitre 5.

Chapitre 5: Classification à large échelle de documents biomédicaux

1 Introduction

Dans la littérature, des travaux majeurs ont été menés sur l'indexation des documents biomédicaux. En ce qui concerne l'indexation conceptuelle, des concepts issus de ressources sémantiques sont utilisés pour représenter les documents. Généralement, on considère que l'ensemble des concepts pertinents permettant de représenter un document sont mentionnés dans ce dernier. Les travaux se focalisent ainsi sur l'identification de ces concepts. Toutefois, des concepts qui peuvent être représentatifs ne sont pas toujours explicitement mentionnés, surtout dans le domaine de la littérature scientifique biomédicale où il n'est pas rare que seule une partie des documents (titre, résumé d'articles scientifiques) soit librement accessible pour des questions de droits d'auteur. Dans ce genre de situation, se limiter à identifier les concepts pertinents uniquement sur la partie disponible des documents à traiter ne permet pas leur représentation complète. Ainsi, arriver à caractériser, à l'aide de concepts issus d'une ressource sémantique, un document entier à partir d'une portion incomplète de textes est un enjeu majeur (Tsoumakas et al., 2010).

D'autre part, l'indexation de documents biomédicaux où chaque document est indexé par un ou plusieurs concepts (appelés aussi catégories ou labels) peut être apparentée à la problématique de classification multi-label. Cette dernière a été largement explorée, notamment dans le cadre de la classification textuelle. Les méthodes proposées peuvent être réparties en deux catégories (Tsoumakas et al., 2010) : l'approche basée sur la transformation du problème en multiples sous-problèmes et l'approche d'adaptation d'algorithmes existants. La première décompose le problème de classification multi-label en un ensemble de problèmes de classification binaire (Papanikolaou et al., 2014). La deuxième adapte les méthodes d'apprentissage existantes pour prendre en compte la classification multi-label (Huang et al., 2011). Des chercheurs ont également investigué la combinaison de ces deux approches (Liu et al., 2014).

Grâce à leur simplicité et leur efficacité en termes de complexité temporelle, de coût d'exécution et de performance, les approches basées sur l'algorithme des k plus proches voisins (*k-nearest neighbor algorithm*, ou tout simplement *k-NN*) ont été largement utilisées (Spyromitros et al., 2008; Zhang et Zhou, 2007; Huang et al., 2011) dans le contexte de la classification multi-label. De même, l'analyse sémantique explicite (*explicit semantic analysis* ou *ESA*) (Gabrilovich et Markovitch, 2007) s'est montrée prometteuse pour la représentation sémantique des textes. L'ESA représente les documents textuels dans un espace conceptuel de grande dimension constitué de concepts extraits à partir de la base de connaissances Wikipédia.

L'objectif du travail présenté dans ce chapitre est de proposer une approche pour décrire automatiquement, à l'aide de descripteurs sémantiques, des documents textuels issus d'un large corpus, quand seule une information partielle sur ces documents est disponible. Pour

cela, nous suivons une voie basée sur la classification automatique supervisée. Nous proposons ainsi trois stratégies de classification. La première, appelée KNN-Classifier, combine la méthode des k plus proches voisins et des méthodes classiques d'apprentissage automatique (Naives Bayes (John et Langley, 1995) ou Random Forest (Breiman, 2001)). La deuxième, nommée ESA-Classifier, est basée sur l'ASE (Gabrilovich et Markovitch, 2007) tandis que la troisième stratégie, appelée Bi-Classifier, combine les deux précédentes. Ces différentes stratégies sont évaluées sur une large collection standard de test extraite de la base MEDLINE.

La suite du chapitre est organisée comme suit. Nous présentons la stratégie KNN-Classifier basée sur l'algorithme des k plus proches voisins dans la section 2. Ensuite, dans la section 3, nous décrivons la stratégie ESA-Classifier qui repose sur l'ASE. La section 4 présente la stratégie consistant à combiner les deux premières : Hybride-Classifier. La section 5 s'intéresse à l'évaluation de ces stratégies sur une large collection de test standard. Nous terminons par une conclusion et des perspectives en section 5.

2 KNN-Classifier : classification basée sur les k plus proches voisins

L'algorithme des k plus proches voisins fait partie de la famille des algorithmes d'apprentissage basés sur les instances[69] (instance-based learning) (Aha et al., 1991). Concrètement, pour prédire la classe (classification) ou la valeur (régression) d'une nouvelle instance, cette dernière est comparée aux instances stockées dans l'ensemble d'entrainement. Pour ce faire, une fonction de similarité est utilisée afin de calculer la distance entre deux instances données.

Dans le cadre de ce travail, le principe est de considérer, pour classifier un document, les concepts assignés manuellement aux documents les plus proches (donc jugés similaires) de ce dernier. Ensuite, le score de pertinence pour représenter le document est estimé pour chacun de ces concepts candidats et les Top concepts les plus pertinents sont retenus.

L'approche que nous proposons comprend ainsi deux étapes. D'abord, pour un document donné, représenté par un vecteur de termes, l'ensemble des documents qui lui sont le plus similaire sont retrouvés (section 2.1). Pour cela, le schéma de pondération TF.IDF a été utilisé afin de déterminer les poids des différents termes dans les documents. Ensuite, la mesure du cosinus a permis d'estimer la similarité entre les documents. Une fois les documents les plus proches du document déterminés, l'ensemble des concepts assignés à ces derniers forment les candidats pour l'annoter. Par la suite, des algorithmes d'apprentissage automatique sont utilisés afin d'ordonner et sélectionner les concepts les plus pertinents pour représenter le document (section 2.2). Pour le filtrage des concepts, nous avons exploré différents algorithmes d'apprentissage et des attributs variés utilisés pour représenter les instances à classer.

[69] La classification (respectivement la régression) consiste à catégoriser (respectivement à prédire la valeur) les objets en fonction de leurs propriétés, appelées aussi attributs. Une instance est définie par une suite de valeurs d'attributs.

2.1 Recherche des documents voisins

KNN-Classifier nécessite une collection de documents préalablement annotés constituant l'espace pour la recherche des documents voisins. Pour un document donné, l'objectif est de retrouver les k documents les plus proches sémantiquement de ce dernier. Pour cela, à l'instar de l'approche *PubMed Related Citations* (Lin et Wilbur, 2007) qui considère que deux documents sont similaires s'ils traitent les mêmes thèmes, nous estimons la similarité entre deux documents en nous basant sur leur contenu. Notre méthode exploite ainsi les termes communs entre les documents pour estimer leur similarité. La mesure du cosinus permet de déterminer par la suite le degré de similarité d'un document donné par rapport aux autres documents de l'espace de recherche. Cette mesure est couramment utilisée dans la classification de textes et en RI avec le modèle vectoriel (Salton et al., 1975) (cf section 3.2 du chapitre 2).

2.1.1 Prétraitement des documents et construction des vecteurs

Les documents sont d'abord segmentés en phrases et en tokens et les mots vides sont supprimés. A partir de ces textes prétraités, tous les uni-grammes et bi-grammes sont extraits et normalisés en utilisant une technique de *stemming*. Ces termes avec leurs poids associés permettent de construire les vecteurs représentant les documents.

2.1.2 Calcul de la similarité entre documents

Une fois tous les documents représentés dans cet espace vectoriel, le cosinus détermine les k documents les plus similaires à un document donné. Pour le calcul des poids, nous utilisons comme indiqué le schéma TF.IDF. Formellement, soit $C = \{d_1, d_2, \dots, d_n\}$, une collection de n documents, $T = \{t_1, t_2, \dots, t_m\}$, l'ensemble des termes apparaissant dans les documents de la collection ainsi que les documents d_i et d_j représentés respectivement par les vecteurs pondérés :

$d_i = (w_1^i, w_2^i, \dots, w_m^i)$ et $d_j = (w_1^j, w_2^j, \dots, w_m^j)$

Leur similarité est définie par :

$$Sim\left(d_i, d_j\right) = \frac{\sum_{k=1}^{m} w_k^i w_k^j}{\sqrt{\sum_k^m (w_k^i)^2}\sqrt{\sum_k^m (w_k^j)^2}}$$

avec w_k^l, le poids du terme t_k dans le document d_l. Il correspond ici à la valeur TF.IDF du terme.

Pour optimiser la recherche, les documents contenus dans l'espace de recherche sont indexés au préalable en utilisant la bibliothèque open source d'indexation et de RI Apache Lucene[70] (McCandless et al., 2010). La recherche des k plus proches voisins devient ainsi un problème de RI où le document cible constitue la requête à traiter. La figure 24 illustre ce processus de recherche de documents proches.

[70] http://lucene.apache.org/core/

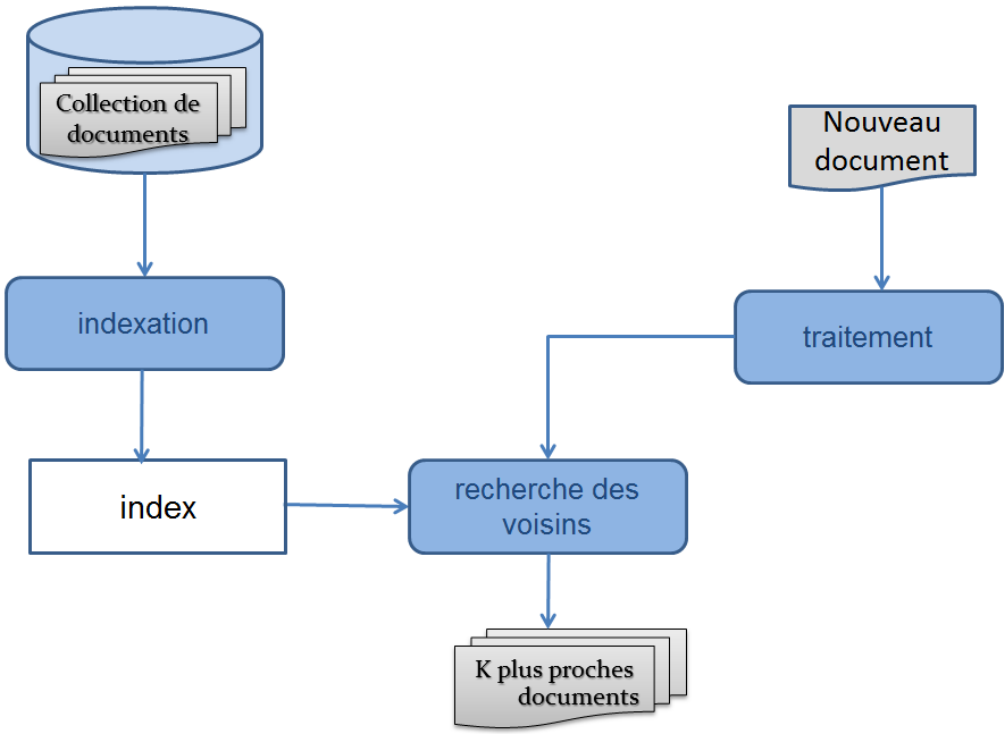


Figure 24 : Processus de recherche des k plus proches voisins

2.2 Classification des documents avec KNN-Classifier

Pour un document cible donné, une fois ses k plus proches voisins retrouvés, tous les labels (concepts) assignés à ces documents sont collectés pour constituer l'ensemble des labels candidats susceptibles d'annoter le document cible. Puisque ceci peut être apparenté à un problème de classification, nous proposons d'utiliser les techniques d'apprentissage automatique pour classifier ces labels ; des algorithmes de classification classiques sont ainsi utilisés pour déterminer les labels pertinents qui vont annoter le document à partir de cet ensemble de candidats. Ainsi, pour chaque label candidat, sa pertinence pour le document cible est prédite. Ensuite, les candidats sont classés en fonction de leur pertinence et les N labels les plus pertinents pour le document sont sélectionnés, N étant fixé empiriquement. Nous avons exploré différentes techniques pour déterminer la valeur optimale de N. Pour l'implémentation de KNN-Classifier, nous avons utilisé l'outil Weka (Waikato Environment for Knowledge Analysis)[71] qui intègre un nombre important d'algorithmes d'apprentissage automatique (Hall et al., 2009). En plus d'être open source, Weka peut être intégré facilement dans un programme Java. Il est également bien adapté pour le développement de nouveaux algorithmes d'apprentissage.

2.2.1 Sélection de la valeur optimale de N

a) Dans un premier temps, N a été fixé comme étant le nombre de labels ayant un score de pertinence supérieur ou égal à un seuil fixé arbitrairement à 0,5. Cette technique qui exploite

[71] http://www.cs.waikato.ac.nz/ml/weka/

seulement le score de pertinence du label par rapport au document est inspirée de la méthode originale des k-NN.

b) Nous avons ensuite fixé la valeur de N comme étant la taille (nombre de labels qui lui sont assignés) moyenne des ensembles de labels collectés parmi les voisins. Cette technique a déjà été utilisée dans une extension de la méthode des k-NN proposée dans (Spyromitros et al., 2008).

c) Enfin, nous avons utilisé la méthode présentée dans (Mao et Lu, 2013). L'idée est de comparer les scores des labels successifs dans la liste de labels candidats classés par ordre décroissant pour fixer une condition d'arrêt. Cette technique, que nous nommons ***règle 6***, est définie comme suit :

$$\frac{s_{i+1}}{s_i} \geq \frac{i}{i+1+\alpha}$$

avec s_i le score de pertinence du label se trouvant à la position i et α une constante dont la valeur optimale est fixée empiriquement.

L'idée de cette règle est de considérer un label comme pertinent tant que le rapport de son score de pertinence par rapport à celui de son prédécesseur dépasse une valeur limite variable en fonction de sa position au sein de la liste des candidats.

2.2.2 Entrainement des classifieurs et classification de nouveaux documents

Pour entrainer les classifieurs, un ensemble d'entrainement consistant en une collection de documents avec les labels qui leur sont assignés est créé en amont ; ces documents sont généralement annotés manuellement. Pour chaque document de cet ensemble d'entrainement, ses k plus proches voisins sont retrouvés et leurs labels associés sont collectés. Chaque label de cet ensemble collecté constitue ainsi une instance de l'ensemble d'entrainement. Ensuite, cet ensemble est utilisé pour construire un modèle de classification. Pour cela, différentes méthodes de classification sont investiguées : le modèle bayésien (Naive Bayes ou NB) (John et Langley, 1995) et les forêts aléatoires (Random Forest ou RF) (Breiman, 2001).

Pour classer un nouveau document, les labels collectés à partir de ses voisins sont représentés comme ceux de l'ensemble d'entrainement. Ensuite, le modèle construit précédemment permet d'estimer le score de pertinence correspondant à chacun de ces labels. En effet, pour chaque label, le modèle calcule sa probabilité d'être pertinent et non pertinent pour annoter le document. Ces probabilités permettent de déterminer le score de pertinence pour chaque label et, par conséquent, de classer l'ensemble des labels candidats en fonction de leurs scores.

2.3 Extraction des attributs

Pour déterminer sa pertinence via un modèle d'apprentissage, chaque label doit être représenté par un vecteur d'attributs. Ces derniers sont les variables sur lesquelles s'appuie le modèle pour prédire la classe à laquelle devrait appartenir le label. Dans la phase d'entrainement où les documents sont préalablement annotés, cette classe prend la valeur 1 si le label est assigné au document cible, sinon elle prend la valeur 0. Dans la phase de

prédiction, le modèle utilise les attributs du label pour estimer son score de pertinence. Nous avons défini six attributs en nous basant sur une synthèse de ce qui a été jugé pertinent dans la littérature et en proposant d'attribuer de l'importance aux labels présents dans les titres des documents (Tableau 14).

- Attribut 1 : pour chaque label candidat, le nombre de documents voisins auxquels il est assigné est utilisé comme attribut. Cette valeur représente un indice important pour déterminer la classe du label. D'ailleurs, dans l'approche k-NN classique, c'est le seul facteur servant à classifier une nouvelle instance. En pratique, une technique de vote permet d'assigner l'instance à la classe qui est la plus répandue parmi ses k plus proches voisins.

- Attribut 2 : Les scores de similarité entre le document à annoter et ses k plus proches voisins annotés avec un label candidat sont additionnés et cette somme constitue également un autre attribut pour ce label.

Puisque la distance entre un document et chacun de ses voisins n'est pas la même, nous considérons que la pertinence des labels assignés à ces derniers pour annoter le document cible est inversement proportionnelle à cette distance. Autrement dit, plus un document est proche du document cible, plus ses labels associés sont susceptibles d'être pertinents pour ce dernier. Dans (Trieschnigg et al., 2009), c'est ce seul attribut qui est utilisé pour déterminer les scores de pertinence des labels candidats. Formellement, comme définis dans (Spyromitros et al., 2008), soient $L = \{l_j\}, j = 1, \dots, n$, l'ensemble des labels candidats pour un nouveau document d et $V = \{d_i\}, i = 1, \dots, k$, l'ensemble de ses voisins les plus proches. Les valeurs de ces deux attributs pour le label l_j sont respectivement définies par :

$$f_1(l_j) = \frac{1}{k}\sum_{i=1}^{n} assigné(l_j, d_i)$$

$$f_2(l_j) = \frac{1}{k}\sum_{l_j \in d_i}^{n} sim(d, d_i)$$

où la fonction binaire $assigné(l_j, d_i)$ renvoie 1 si le label l_j est assigné au document d_i, 0 sinon ; et $sim(d, d_i)$ est le score de similarité entre les documents d et d_i calculé en utilisant la mesure du cosinus décrite dans la section 3.2 du chapitre 2.

- Attribut 3 : pour chaque label candidat, nous avons également vérifié si tous les tokens le constituant apparaissaient dans le document. Cet attribut binaire permet de capturer de manière simpliste les termes constitués de mots disjoints, qui sont fréquents dans les textes biomédicaux.

Par ailleurs, nous avons calculé deux attributs en nous basant sur les termes synonymes.

Pour l'indexation des documents biomédicaux, le thésaurus MeSH est généralement utilisé. Comme déjà dit dans le chapitre 1 (section 6.1), ce dernier est composé d'un ensemble de descripteurs (appelés aussi *Main headings*) organisés selon une structure hiérarchique. Chaque descripteur comprend des termes synonymes et des termes associés, qui constituent ses entrées. Ainsi, les attributs 4 et 5 sont définis comme suit :

- Attributs 4 : pour chaque descripteur, si une de ses entrées (synonymes et termes associés) apparait dans le document, le quatrième attribut prend la valeur 1.
- Attributs 5 : la fréquence du descripteur dans le document considéré est affectée au cinquième attribut.

Ces deux attributs se voient attribués la valeur nulle si aucune de ses entrées (synonymes et termes associés) n'apparait dans le document,

- Attribut 6 : cet attribut binaire prend la valeur 1 lorsque le label candidat est contenu dans le titre du document. Nous supposons logiquement que si un label apparait dans le titre d'un document, ceci renforce son importance pour représenter ce document.

Tableau 14 : Les différents attributs utilisés pour entrainer les classifieurs

Attribut	Description
Attribut 1	nombre de documents voisins auxquels le label est assigné
Attribut 2	somme des scores de similarité entre le document et les voisins dans lesquels le label apparait
Attribut 3	présence de tous les tokens constituant du label candidat (à l'exception des mots vides) dans le document
Attribut 4	présence d'une des entrées du label candidat dans le document
Attribut 5	fréquence du label dans le document
Attribut 6	présence du label dans le titre du document

Le tableau 15 montre des exemples de documents similaires à un document et les labels utilisés pour les annoter.

Tableau 15 : Les documents les plus proches du document ayant pour PMID 23192094 avec les labels utilisés pour les annoter

22353656	Earthquakes - Radioactive Hazard Release - Diffusion - Tsunamis - Soil Pollutants, Radioactive - Japan - Nuclear Power Plants -
22428463	Earthquakes - Questionnaires - Radioactive Hazard Release - Humans - Adult - Middle Aged - Public Opinion - Male - Japan - Female - Nuclear Power Plants -
23439139	Earthquakes - Radioactive Pollutants - Radioactive Hazard Release - Luminescent Measurements - Humans - Environmental Exposure - Radiographic Image Enhancement - Radiation Monitoring - Japan - Nuclear Power Plants -
23842513	Mental Disorders - Aged, 80 and over - Humans - Adult - Fukushima Nuclear Accident - Aged - Middle Aged - Inpatients - Adolescent - Male - Japan - Female -

22955043	Earthquakes - Humans - Disasters - Fukushima Nuclear Accident - Mental Health - Child - Dose-Response Relationship, Radiation - Thyroid Gland - Pregnancy - Child, Preschool - Infant - Health Surveys - Follow-Up Studies - Adolescent - Radiation Monitoring - Female - Japan -
22955043	Earthquakes - Humans - Disasters - Fukushima Nuclear Accident - Mental Health - Child - Dose-Response Relationship, Radiation - Thyroid Gland - Pregnancy - Child, Preschool - Infant - Health Surveys - Follow-Up Studies - Adolescent - Radiation Monitoring - Female - Japan -
22549322	Soil - Radiation Dosage - Radioactive Hazard Release - Radioisotopes - Radioactive Fallout - Environmental Exposure - Data Collection - Time Factors - Radiation Monitoring - Japan - Nuclear Power Plants -
21799088	Thyroid Neoplasms - Environmental Health - Radioactive Hazard Release - Uranium - Environmental Exposure - Disasters - Tsunamis - Nuclear Reactors - Potassium Iodide - Japan - Nuclear Power Plants -
23982606	Radioactive Pollutants - Vegetables - Germanium - Food Contamination - Iodine Radioisotopes - Fukushima Nuclear Accident - Decontamination - Cesium Radioisotopes - Autoradiography - Radiation Monitoring - Nuclear Power Plants -
22353655	Earthquakes - Radiation Dosage - Gamma Rays - Radioactive Hazard Release - Humans - Tsunamis - Radiation Monitoring - Radon - Japan - Nuclear Power Plants -
22469934	United States - Earthquakes - Radiation Dosage - United States Environmental Protection Agency - Public Health - Radioactive Hazard Release - Emergencies - Tsunamis - Radiation Monitoring - Japan - Nuclear Power Plants -
22378205	Environment - Radioactive Hazard Release - Radioisotopes - Radiation Monitoring - Japan - Nuclear Power Plants -
22845725	Occupational Exposure - Radiation Dosage - Occupational Health - Radioactive Hazard Release - Humans - Disasters - Decontamination - Cesium Radioisotopes - Dust - Nuclear Power Plants - Air Pollutants, Radioactive - Radiation Monitoring - Radi ation Injur ie s - Fem al e - He alth Priorities - Japan - Male -
23642080	Radiation Dosage - Gamma Rays - Humans - Adult - Fukushima Nuclear Accident - Air - Whole-Body Counting - Thyroid Gland - Time Factors - Male - Female - Nuclear Power Plants -
23982615	Soil - Occupational Exposure - Young Adult - Medical Records - Humans - Adult - Fukushima Nuclear Accident - Decontamination - Middle Aged - Radiation Monitoring - Male - Nuclear Power Plants -
22864411	Swine - Animals - Blood Chemical Analysis - Biological Assay - Radioactive Fallout - Fukushima Nuclear Accident - Cesium Radioisotopes - Soil Pollutants, Radioactive - Iodine Radioisotopes - Feces - Radiation Monitoring - Japan - Male -
23952577	Occupational Exposure - Radiation Dosage - Humans - Air Pollutants, Radioactive - Adult - Radioisotopes - Fukushima Nuclear Accident - Aged - Middle Aged - Japan -
22469931	Earthquakes - Radiation Dosage - Radiometry - Radioactive Hazard Release - Background Radiation - Tsunamis - Japan - Nuclear Power Plants -
22059981	Earthquakes - Environment - Occupational Exposure - Radiation Dosage - Radioactive Hazard Release - Information Dissemination - Water Supply - Temperature - Disasters - Hydrogen - Nuclear Reactors - Explosions - Nuclear Power Plants - Cities - Radiologic H ealth - R adiois otopes - Tsunamis - Time Factors - Radiation Monitoring - Japan -
23274827	Radiochemistry - Inhalation - Radioactive Pollutants - Radiation Protection - Protective Clothing - Half-Life - Rhinitis, Allergic, Seasonal - Humans - Environmental Exposure - Fukushima Nuclear Accident - Masks -

A partir de l'ensemble des labels associés à ces documents, un classifieur est utilisé pour estimer la pertinence de chaque label pour le document cible. Pour cela, leurs différents

attributs sont préalablement discrétisés. Ensuite, les classifieurs tels que Naive Bayes et Random Forest sont appliqués.

Nous allons à présent décrire la seconde stratégie de classification, ESA-Classifier, qui utilise des mesures statistiques simples pour établir les associations entre les mots et les concepts.

3 ESA-Classifier : classification d'une large collection de documents en utilisant l'analyse sémantique explicite

Dans cette section, nous présentons d'abord l'ASE. Ensuite, nous décrivons en détail notre stratégie de classification de documents biomédicaux basée sur cette technique.

3.1 L'analyse sémantique explicite

L'ESA est une approche similaire à l'analyse sémantique latente (LSA), proposée pour la représentation sémantique de documents textuels (Gabrilovich et Markovitch, 2006; Gabrilovich et Markovitch, 2007). Dans cette méthode, les documents sont représentés dans un espace conceptuel de grande dimension constitué de concepts explicites extraits automatiquement à partir de la base de connaissances Wikipédia. Pour cela, des techniques statistiques sont utilisées afin de représenter explicitement n'importe quel texte (mots simples, fragments de texte, document entier) par des vecteurs pondérés de concepts Wikipédia. Dans cette approche, les titres des articles de Wikipédia sont définis comme étant les concepts. Ainsi, chaque concept est représenté par un vecteur constitué de l'ensemble des mots (suppression éventuelle des mots vides) qui apparaissent dans l'article Wikipédia correspondant. Les poids associés à ces mots, représentant les entrées du vecteur, sont les scores d'association entre ces derniers et le concept. Ces poids sont calculés en utilisant le schéma de pondération TF.IDF.

A partir de cette représentation, chaque concept (titre de Wikipédia) est représenté par un vecteur pondéré de mots. Ensuite, un index inversé, dans lequel chaque mot est représenté par un vecteur de concepts associés, est créé. Dans cet index, les concepts moins importants (i.e., ayant un poids faible) pour un vecteur sont éliminés. L'index est, par la suite, utilisé pour classifier les documents textuels.

Le processus de classification comprend deux étapes. Un document donné est d'abord représenté par un vecteur de mots en utilisant la mesure TF.IDF. Les concepts correspondant à ces mots sont ensuite retrouvés dans l'index inversé et fusionnés pour constituer un vecteur de concepts représentant le document. Les concepts retrouvés sont classés par ordre décroissant de leur score de pertinence par rapport au document. Ce processus est illustré par la figure 25.

Formellement, considérons un texte T, $\{w_i\}$ l'ensemble des mots apparaissant dans T et $\langle v_i \rangle$ leur poids respectif. Soit $\langle k_j \rangle$ le score d'association entre w_i et le concept c_j avec $c_j \in C$ où C est l'ensemble des concepts de Wikipédia. Le poids (ou encore la pertinence) d'un concept, pour le texte T, est défini par :

$$poids\ (c_j) = \sum_{w_i \in T} v_i . k_j$$

Dans leurs expérimentations sur le calcul de proximité sémantique entre textes, Gabrilovich et Markovitch ont rapporté de bonnes performances de l'ESA. En pratique, la corrélation avec les jugements humains est passée de 0,56 (état de l'art) à 0,75 pour des mots simples et de 0,60 à 0,72 pour des textes (Gabrilovich et Markovitch, 2007).

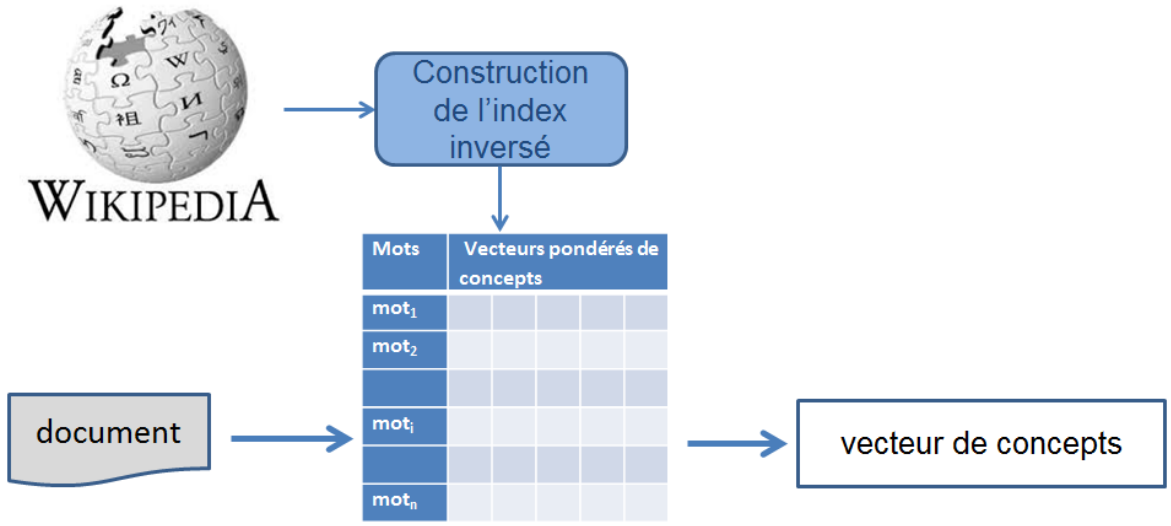


Figure 25 : Processus de l'analyse sémantique explicite

Compte tenu de ces résultats prometteurs, il nous a paru intéressant d'investiguer cette approche pour la classification de larges collections de documents biomédicaux.

3.2 La stratégie ESA-Classifier basée sur l'analyse sémantique explicite

Tout d'abord, comme hypothèse de départ, nous supposons qu'une ressource sémantique contenant les concepts à utiliser pour classifier les documents est disponible ainsi qu'un ensemble d'entrainement où chaque document est annoté par un ensemble de concepts. Contrairement à la méthode ESA originale où chaque article de Wikipédia est associé à un seul concept, dans notre approche, chaque document de l'ensemble d'entrainement peut être annoté par un ou plusieurs concepts.

A partir de l'ensemble de documents d'entrainement, nous utilisons des techniques statistiques pour établir des associations entre les concepts et les mots extraits dans les textes. Ainsi, pour chaque concept, les mots qui lui sont plus fortement associés sont utilisés pour le représenter. Si les concepts sont vus comme des documents, on se retrouve confronté à un problème de RI où l'objectif est de trouver les documents (concepts) les plus pertinents par rapport à une requête donnée (un nouveau document). Par conséquent, les modèles classiques de RI peuvent être utilisés pour représenter les documents et les requêtes, mais également pour calculer la pertinence d'un document par rapport à une requête. Ces derniers permettent donc de classer les concepts retournés par ordre décroissant de leur pertinence pour un

document donné. Enfin, les concepts les plus pertinents sont sélectionnés pour représenter un document.

Comme dans la méthode basée sur les k-NN, les documents sont traités en utilisant les mêmes techniques : segmentation en phrases, tokenisation, suppression des mots vides et normalisation en utilisant l'algorithme de stemming de Porter (Porter, 1980). Toutefois, dans cette approche, seuls les uni-grammes sont utilisés pour la représentation des documents. Ce choix vise à limiter la taille des vecteurs de termes représentant les concepts et à simplifier ainsi le calcul.

Dans (Sorg et Cimiano, 2012), les auteurs ont montré, à travers leurs expérimentations, que l'approche ESA est très sensible à certains facteurs, tels que la fonction utilisée pour le calcul des scores d'association entre les concepts et les mots. Nous avons ainsi étudié différentes mesures pour estimer le score d'association optimal entre les concepts et les mots, mais également pour déterminer la pertinence d'un concept pour représenter un document donné. Pour le calcul des scores d'association entre un concept c et un mot w, nous avons expérimenté les mesures suivantes :

- la fonction **TF.ICF** (la mesure TF.IDF adaptée aux concepts) (Salton et al., 1975) :

$$TF.ICF(w,c) = TF(w,c) * \log\frac{N}{n_i}$$

avec N le nombre total de concepts, n_i le nombre de concepts associés au mot w. Le facteur $TF(w,c)$ est le nombre d'occurrences de w dans des documents annotés par le concept c normalisé et est défini par :

$$TF(w,c) = \sum_{d \in D_c} \frac{freq(w,d)}{|d|}$$

où $freq(w,d)$ est la fréquence du mot w dans le document d, $|d|$ la taille en termes de nombre de mots de d et D_c l'ensemble des documents annotés par le concept c.

- l'**indice de Jaccard** (Jaccard, 1912) **:**

$$J(w,c) = \frac{cocc(w,c)}{occ(w) + occ(c) - cocc(w,c)}$$

avec $cocc(w,c)$ le nombre de documents où co-occurrent le concept c et le mot w, $occ(c)$ le nombre de documents annotés par le concept c et $occ(w)$ le nombre de documents où apparait le mot w.

Ensuite, pour estimer la pertinence d'un concept susceptible d'annoter un document, nous avons appliqué la mesure ci-après. Ainsi, le score de pertinence d'un concept c pour un nouveau document d est défini par :

$$Pert(c,d) = \sum_{w \in d} WF(w,d) * score(w,c)$$

avec $score(w,c)$ le score d'association entre le mot w et le concept c et $WF(w,d)$ la fréquence du mot w dans le document d.

Après avoir détaillé les deux stratégies de classification de documents, nous décrivons brièvement la stratégie hybride qui combine celles-ci.

4 Stratégie hybride de classification d'une large collection de documents

L'idée est d'intégrer les informations obtenues avec la stratégie basée sur l'ESA et celles issues de la recherche des k-NN dans un seul modèle de classification. Pour cela, en plus des attributs utilisés précédemment dans notre méthode de classification basée sur les k-NN, le score de pertinence généré avec la technique ESA est considéré comme un attribut supplémentaire. Le principe est le suivant :

a) Pour chaque document, les labels qui lui sont le plus fortement associés sont retrouvés avec leur score de pertinence correspondant en utilisant la méthode ESA.

b) Ensuite, pour chaque label candidat retrouvé avec les k plus proches voisins, son nouvel attribut prend pour valeur son score de pertinence s'il est parmi ces labels préalablement trouvés, 0 sinon.

c) Enfin, de nouveaux modèles, entrainés en utilisant tous ces attributs et les mêmes classifieurs, permettent de prédire les labels pertinents pour annoter un nouveau document.

5 Evaluation des différentes stratégies

Pour évaluer les performances de nos méthodes, nous avons réalisé deux expérimentations : une première expérimentation avec des collections de données fournies par les organisateurs du challenge international BioASQ[72] (Balikas et al., 2014) et une seconde expérimentation avec une collection constituée d'extraits les plus récents (publiés depuis 2013).

5.1 Les collections de données

5.1.1 Jeu de données de BioASQ

BioASQ (Tsatsaronis et al., 2012) est une campagne d'évaluation qui vise à promouvoir le développement de systèmes permettant de faciliter l'accès à l'information biomédicale. L'enjeu est de pouvoir fournir des informations spécifiques aux experts à partir de larges sources de connaissances parfois hétérogènes et de bases de données. Elle est constituée de deux tâches : une tâche qui s'intéresse à l'indexation sémantique de vastes quantités d'informations (tâche a, tâche *1*a pour l'année 2013 et *2a* pour 2014) et une autre focalisée sur les systèmes de questions/réponses (*question/answering*) permettant d'interpréter les

[72] http://bioasq.lip6.fr/

questions des utilisateurs et de retourner des réponses appropriées (tâche b). Chaque tâche comprend plusieurs batchs et chaque batch regroupe un ensemble de tests (5 pour la tâche 2a). Nous avons participé à la tâche 2a qui a pour objectif l'indexation avec des descripteurs du thésaurus MeSH d'une large collection d'articles dont seule une partie du texte est disponible.

Les organisateurs du challenge, dans son édition 2014, ont fourni une collection de plus de quatre millions de documents issus de revues spécifiques du domaine biomédical, extraits de la base de données MEDLINE. Cette collection était ainsi constituée uniquement des titres et résumés d'articles scientifiques extraits à partir de la base bibliographique PubMed et annotés manuellement. Durant le challenge, les organisateurs fournissaient, chaque semaine, des articles de PubMed non encore annotés, considérés comme des collections de test pour évaluer les systèmes participant à la tâche 2a. Les participants devaient classifier ces ensembles de test en utilisant le thésaurus MeSH. Ces ensembles de test ont, par la suite, été annotés par des indexeurs humains de PubMed pour évaluer les résultats fournis par les différents systèmes participant au challenge.

Pour la recherche de documents similaires, nous avons considéré, dans notre première phase d'expérimentations, tous les articles de cette collection publiés depuis 2000. L'idée était d'ignorer les articles anciens pour ne pas pénaliser les descripteurs ajoutés récemment dans le thésaurus MeSH. Cet espace de recherche a été ensuite étendu à l'ensemble de la collection.

5.1.2 Jeu de données extrait de la collection de BioASQ

Pour la seconde expérimentation, nous avons extrait, à partir de la collection précédente, l'ensemble des articles publiés depuis 2013 parmi lesquels 20 000, sélectionnés aléatoirement, ont été utilisés comme ensemble d'entrainement pour les classifieurs et 1000 autres comme ensemble de tests. Ces données permettant d'entrainer les classifieurs ont été ensuite étendues jusqu'à 50 000 documents, dans le but d'améliorer les performances de la classification. La collection de test a également été augmentée à 2 000 documents. Comme dans les données d'entrainement, chaque document dans la collection de test est annoté manuellement par un ensemble de descripteurs afin d'évaluer les résultats de nos différentes méthodes.

La même collection d'entrainement a été également utilisée lors des deux expérimentations pour entrainer les classifieurs.

Concernant notre deuxième stratégie basée sur l'ESA, hormis les documents de l'ensemble de test qui ont été utilisés pour l'évaluation, tout le reste de la collection (soit 4 430 399 documents) a été exploité afin de déterminer les associations entre les mots et les différents concepts de la ressource.

5.2 Les mesures d'évaluation

Rappelons que l'indexation de documents biomédicaux peut être apparentée à un problème de classification multi-label. Au lieu d'une seule classe, chaque document est indexé par un ensemble de labels. Ainsi, pour l'évaluation, les organisateurs de BioASQ proposent d'utiliser un ensemble de mesures adaptées parmi lesquelles nous avons sélectionné les suivantes : a) la précision (**EBP** ou example based precision), b) le rappel (**EBR** ou example based recall), c)

la f-mesure (**EBF** ou example based f-measure), et d) l'exactitude (**Acc** ou Accuracy). Ces mesures sont calculées comme suit. Considérons Y_i, les labels assignés manuellement aux documents, Z_i l'ensemble des labels prédits par un système et m le nombre de documents de l'ensemble de test, alors ces différentes mesures sont définies comme suit :

$$EBP = \frac{1}{m}\sum_{i=1}^{m}\frac{|Yi \cap Zi|}{|Zi|}$$

$$EBR = \frac{1}{m}\sum_{i=1}^{m}\frac{|Yi \cap Zi|}{|Yi|}$$

$$EBF = \frac{1}{m}\sum_{i=1}^{m}\frac{|Yi \cap Zi|}{|Zi|+|Yi|}$$

$$Acc = \frac{1}{m}\sum_{i=1}^{m}\frac{|Yi \cap Zi|}{|Y_i \cup Zi|}$$

Ces différentes mesures, en plus d'être courantes, sont représentatives et permettent de juger globalement les performances des systèmes.

5.3 L'environnement d'évaluation

Dans nos différentes expérimentations, nous avons utilisé le cluster du Mesocentre de Bordeaux, Avakas[73] qui comprend :

- les nœuds de calcul c6100 (x264): ce sont les machines sur lesquelles les travaux sont exécutés. Ils ont les caractéristiques suivantes :
    - deux processeurs hexacœurs (12 cœurs par nœud) Intel® Xeon® x5675 @ 3,06 GHz
    - 48 Go RAM

- les nœuds de calcul bigmem R910 (x4) : ces nœuds qui ont plus de mémoire et de cœurs ont des processeurs plus lents :
    - processeurs 10 cœurs (40 cœurs par nœud) Intel® Xeon® E7-4870 @ 2,4 GHz
    - 512 Go RAMdisques SAS 10krpm

Dans notre cas, nous avons utilisé deux nœuds de calcul c6100, ce qui fournit une mémoire RAM de 48GB et 24 cœurs Intel® Xeon® x5675.

Pour l'entrainement sur une collection de 20 000 documents, le temps de calcul est respectivement 3 et 2 minutes pour le RF (Random Forest) et le NB (Naive Bayes) une fois les données représentées dans un format approprié (i.e. ARFF). Sur une collection élargie de 50 000 documents, le temps d'entrainement augmente jusqu'à neuf minutes environ avec le RF contre cinq quatre minutes avec le NB. La préparation des données (recherche des voisins et calcul des valeurs d'attributs) reste toutefois coûteuse en temps (1 heure et 43 minutes). La classification avec ces deux classifieurs pour un ensemble de test de 2000 documents prend environ 7 minutes.

[73] http://www.mcia.univ-bordeaux.fr/index.php?id=45

5.4 Résultats

5.4.1 Résultats de la stratégie KNN-Classifier

Dans un premier temps, nous présentons les résultats de notre participation à la tâche 2a du challenge BioASQ. Nous rapportons les résultats obtenus dans le batch 3 car c'est celui où notre système a été le plus performant. Le tableau 16 montre les résultats de notre système et de ceux qui ont obtenu les meilleures performances dans les différents tests du batch 3. Dans les tests 2 et 5, notre meilleur système était basé sur le classifieur NB et retenait seulement les labels ayant un score de confiance supérieur ou égal à 0,5. Dans les autres tests (tests 1, 3 et 4), notre meilleur système a été celui dans lequel le nombre de labels (N) pour un document était fixé comme étant la moyenne du nombre de labels de ses voisins. Dans la plupart des cas, les systèmes utilisant cette moyenne pour déterminer N ont donné des résultats meilleurs ou comparables aux autres. Lors de notre participation au challenge, nous n'avons pas testé la méthode décrite dans (Mao et Lu, 2013) pour déterminer la valeur optimale de N.

Tableau 16 : Résultats de notre système comparés à ceux du meilleur système dans les différents tests du batch 3. Taille est le nombre de documents contenus dans le test

Test	Taille	Système	EBP	EBR	EBF
test 1	2 961	KNN-Classifier	0,55	0,48	0,49
		Meilleur système	0,59	0,62	0,58
test 2	5 612	KNN-Classifier	0,52	0,50	0,48
		Meilleur système	0,62	0,60	0,60
test 3	2 698	KNN-Classifier	0,55	0,49	0,49
		Meilleur système	0,64	0,63	0,62
test 4	2 982	KNN-Classifier	0,49	0,55	0,49
		Meilleur système	0,63	0,62	0,62
test 5	2 697	KNN-Classifier	0,50	0,53	0,48
		Meilleur système	0,64	0,61	0,61

Cependant, nous avons exploré cette technique dans la seconde expérimentation. Chaque article de la base MEDLINE est habituellement annoté par 5 à 25 descripteurs MeSH. Dans la collection d'entrainement de la tâche 2a du challenge BioASQ, le nombre moyen de descripteurs utilisés (pour les journaux ciblés) pour indexer un document est de 13,2. Ce nombre est très variable et son estimation a un impact important sur les résultats de la classification.

Dans un deuxième temps, nous avons évalué KNN-Classifier avec différentes configurations sur l'ensemble de test décrit ci-dessus en utilisant le deuxième jeu de données et nous avons comparé les différentes performances obtenues. Ainsi, différents algorithmes d'apprentissage et diverses techniques permettant de fixer le nombre de labels pour annoter un document donné ont été combinés. Les résultats obtenus suivant le classifieur utilisé avec une variation

de N sont présentés dans les tableaux 17 à 19. Le paramètre k a été fixé empiriquement à 20 dans cette expérimentation en utilisant une technique de validation croisée.

Tableau 17 : Résultats de KNN-Classifier en fonction du classifieur utilisé en fixant le seuil minimal du score de confiance à 0,5

Classifieur	EBP	EBR	EBF
NB	0,58	**0,49**	**0,49**
RF	**0,74**	0,34	0,43

Tableau 18 : Résultats de KNN-Classifier en fonction du classifieur utilisé en utilisant la moyenne des nombres de labels des voisins d'un document comme valeur de N

Classifieur	EBP	EBR	EBF
NB	0,51	0,54	0,51
RF	**0,52**	0,54	**0,52**

Tableau 19 : Résultats de KNN-Classifier en fonction du classifieur utilisé en comparant les scores des labels successifs

Classifieur	EBP	EBR	EBF
NB	0,56	0,52	0,51
RF	**0,61**	0,52	**0,53**

Nous remarquons que, lorsque le seuil minimal du score est fixé à 0,5, la précision augmente de manière conséquente, notamment avec le classifieur RF mais le rappel est faible (Tableau Tableau 17). En ce qui concerne la technique basée sur la moyenne du nombre de labels des voisins, elle donne un bon rappel mais la précision diminue légèrement comparativement à la première (Tableau 18). Dans ce cas, les résultats des deux classifieurs sont comparables même si le RF dépasse légèrement le NB. Les meilleurs résultats ont été obtenus en utilisant la technique de comparaison des scores qui maintient un équilibre entre précision et rappel et donne, par conséquent, la meilleure f-mesure. A l'exception de la première technique où les résultats obtenus avec le classifieur NB sont meilleurs (Tableau 17), la meilleure f-mesure est atteinte en utilisant le classifieur RF (Tableaux 18 et 19).

Il convient de noter que nous avons également expérimenté les arbres de décision et les réseaux neuronaux mais les résultats obtenus avec ces derniers sont moins intéressants. Pour le premier, les résultats sont moins bons selon les mesures précédentes, tandis que le second

fournit des résultats comparables à ceux obtenus avec le RF, cependant avec un temps d'exécution très important (de l'ordre de trois fois plus que le RF).

Comparé aux meilleurs systèmes présentés à la tâche 2a du challenge BioASQ 2014, l'ensemble de documents utilisé par notre méthode pour entrainer les classifieurs est restreint (20 000 contre 1 000 000 de documents pour certains systèmes (Balikas et al., 2014)). Nous avons ainsi étendu cet ensemble d'entrainement à 50 000 documents, ce qui a permis d'améliorer significativement les résultats de notre méthode (Tableau 20). La valeur de la constante α, utilisée dans la stratégie de fixation du nombre de labels basée sur la comparaison des scores, impacte également les performances de la classification. La figure 26 illustre la variation des résultats en fonction de α. Plus la valeur de α est petite, plus la précision est élevée mais le rappel est bas et inversement. Dans ces expérimentations, nous avons fixé α à 1,6 puisque c'est cette valeur qui donne les meilleurs résultats. Selon l'exemple présenté dans le tableau 15, une liste de huit labels ont été sélectionnés parmi les 91 candidats annotant ses voisins avec une précision de 0,88 et un rappel de 1. Nous avons comparé les performances de notre approche avec l'annotation manuelle. Le tableau 21 présente cette comparaison pour le document MedLine ayant le PMID *23192094* pour lequel seuls sept concepts avaient été sélectionnés manuellement.

5.4.2 Résultats de la stratégie ESA-Classifier

Du traitement de la collection d'entrainement constituée de 4 432 399 documents (résumés + titres), résultent 1 630 405 mots distincts et 26 631 concepts assignés à ces documents parmi les 27 149 du MeSH 2014 (98,1%). Pour simplifier le calcul et optimiser les résultats de la classification, chaque concept est représenté par un vecteur constitué des 200 termes les plus fortement associés à ce dernier sachant que seuls les termes apparaissant au moins dans cinq documents sont considérés. Ensuite, pour annoter un nouveau document, il est représenté lui aussi par un vecteur de mots et sa similarité par rapport à chaque concept est calculée en utilisant la mesure du cosinus. A l'issue de ce calcul, les concepts les plus similaires au document sont sélectionnés. Ici, nous supposons que le nombre de concepts pour indexer le document est connu (fourni au système) et que, par conséquent, les mesures de précision (EBP) et de rappel (EBR) se valent ; ainsi nous rapportons seulement la f-mesure (EBF) et l'exactitude (Acc).

Après l'évaluation d'ESA-Classifier, nous avons constaté, comme dans des travaux antérieurs, que sa performance varie en fonction de la mesure utilisée pour le calcul du score d'association entre les mots et les concepts. Ce comportement est illustré dans le tableau 22 où la mesure de Jaccard permet d'avoir de meilleurs résultats.

Nous notons aussi que cette méthode, comparée à KNN-Classifier, donne des résultats moins bons. Ceci peut s'expliquer par le fait que, pour classifier les documents, ESA-Classifier se base simplement sur leurs contenus, qui sont partiels dans notre contexte; elle exploite les mots contenus dans les documents et les labels utilisés pour les annoter pour déterminer les scores d'association labels-mots alors que ces documents sont incomplets (ils ne contiennent pas tous les mots représentatifs). En plus, les mots utilisés pour retrouver les labels permettant de classifier un document le représentent partiellement.

Tableau 20 : Résultats de KNN-Classifier avec un ensemble d'entrainement étendu (50 000 documents) et la stratégie présentée dans la règle 6

Classifieur	EBP	EBR	EBF	Acc
NB	0,597	**0,536**	0,54	0,394
RF	**0,628**	0,534	**0,55**	**0,409**

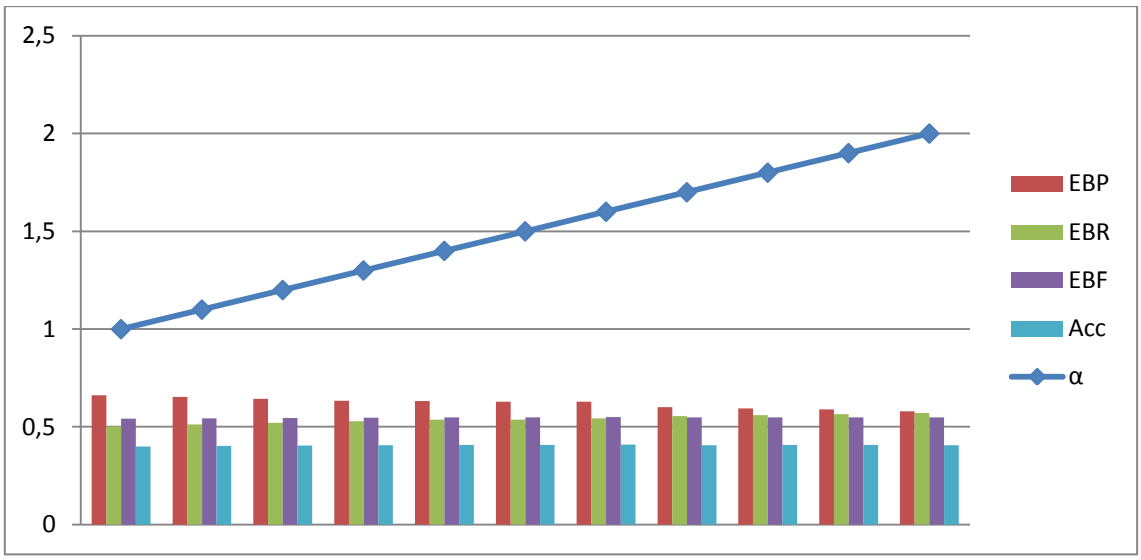


Figure 26 : Variation des performances de KNN-Classifier en fonction de la consante α

Tableau 21 : Liste des huit concepts sélectionnés par KNN-Classifier avec leur pertinence, comparativement à l'annotation manuelle pour le document de PMID 23192094

Labels sélectionnés par KNN-Classifier	Pertinence	Labels sélectionnés manuellement
Earthquakes	0,45	oui
Radiation Dosage	0,39	oui
Radioactive Hazard Release	0,54	oui
Humans	0,57	oui
Nuclear Power Plants	0,84	oui
Radiation Monitoring	0,61	oui
Japan	0,82	oui
Fukushima Nuclear Accident	0,45	non

Tableau 22 : Résultats de ESA-Classifier en fonction du score d'association choisi

Score d'association	EBF	Acc
Indice de Jaccard	0,26	0,16
TF.ICF	0,22	0,13

5.4.3 Résultats de la stratégie mixte Bi-Classifier

A l'issue des expérimentations précédentes, nous avons évalué la stratégie qui consiste à combiner les deux stratégies précédentes de classification afin de voir l'impact sur les performances de notre système. Nous avons noté que cette combinaison n'a pas permis d'améliorer les résultats, contrairement à ce que nous espérions (Tableau 23).

Tableau 23 : Résultats de la combinaison des deux approches

Classifieur	EBP	EBR	EBF	Acc
NB	0,57	**0,54**	0,53	0,38
RF	**0,61**	0,53	**0,54**	**0,40**

5.5 Analyse des résultats

Les expérimentations réalisées en utilisant notre stratégie basée sur les k-NN, KNN-Classifier, montrent que cette dernière est très prometteuse pour la classification de documents biomédicaux. Dans (Trieschnigg et al., 2009), parmi les systèmes présentés, celui basé sur l'approche des k-NN a donné les meilleurs résultats. Comparativement à cette approche, dans notre stratégie, nous utilisons des attributs supplémentaires pour déterminer la pertinence d'un label candidat pour un document donné. En effet, pour estimer le degré de pertinence d'un label pour un document, Trieschnigg et ses collègues additionnent les scores de pertinence entre ce document et ses k plus proches voisins qui sont annotés par ce label. Dans KNN-Classifier, cette valeur constitue simplement un attribut parmi d'autres pour prédire les scores de pertinence des labels. Bien que les résultats de KNN-Classifier ne dépassent pas ceux du système de référence MTI (Mork et al., 2013), utilisé par les indexeurs de la National Library of Medicine, il offre malgré tout des performances encourageantes et comparables (une f-mesure de 0,53 contre 0,56 pour le système MTI), mais qui nécessitent d'être améliorées. Une comparaison directe avec la méthode proposée dans (Huang et al., 2011) n'est pas simple puisque les auteurs ont utilisé une collection de test ancienne différente de celle fournie dans le challenge BioASQ, qui est plus récente et annotée avec la nouvelle version du thésaurus MeSH (version 2014). Toutefois, de manière similaire à leurs expérimentations, lorsque KNN-Classifier est évalué sur une collection de 1000 documents sélectionnés aléatoirement, il obtient des performances meilleures que la méthode et Huang et al. (0,53 de f-mesure contre 0,50). Cependant, dans une comparaison avec leurs résultats récents dans le premier challenge BioASQ (Mao et Lu, 2013) où ils intègrent les sorties du MTI, leur système donne des

résultats meilleurs que ceux obtenus par nos différentes stratégies (une f-mesure de 0,55 contre 0,53). Comparativement à deux autres approches proposées dans (Zhu et al., 2013), l'une basée sur l'outil MetaMap (Aronson et Lang, 2010) et l'autre utilisant des techniques de RI (Zhu et al., 2013), notre stratégie obtient de meilleurs résultats (une f-mesure de 0,53 contre 0,42). Notre approche dépasse également la méthode de classification hiérarchique proposée dans (Ribadas-Pena et al., 2013). Dans le cadre de notre participation au challenge BioASQ, le classifieur NB a été utilisé en fixant N à la moyenne du nombre de labels assignés à ses voisins pour déterminer l'ensemble des labels pertinents permettant d'annoter un document. Dans notre seconde expérimentation, nous avons noté cependant que la combinaison du classifieur RF avec la stratégie basée sur la comparaison des scores des labels successifs a permis d'obtenir de meilleures performances.

Une évaluation plus récente de KNN-Classifier sur une collection d'entrainement plus large (50 000 documents) a montré qu'elle donne de bonnes performances, comparables aux meilleures méthodes décrites dans la littérature. Par ailleurs, contrairement au système MTI, nous n'utilisons pas de règles de filtrage spécifiques, ce qui rend notre approche plus générale et sa réutilisation dans d'autres domaines possible. Notons de plus que les meilleurs systèmes du challenge BioASQ (Liu et al., 2014; Papanikolaou et al., 2014; Mao et al., 2014) utilisent un classifieur binaire pour estimer le score de pertinence de chaque label de la ressource pour chaque document (Balikas et al., 2014), ce qui nécessite beaucoup de ressources en termes de calcul et d'espace de stockage.

Bien qu'elle se soit montrée intéressante, la méthode basée sur l'ESA a donné des résultats très faibles, comparables aux méthodes basiques utilisant une simple correspondance entre les textes et les entrées de la ressource sémantique.

La combinaison de ces deux approches (Bi-Classifier), n'a pas non plus permis d'améliorer les performances de la classification.

6 Conclusion

Dans ce chapitre, nous avons présenté notre approche de classification d'une large collection de textes biomédicaux. Puisque les concepts représentatifs d'un document ne sont pas toujours explicitement mentionnés dans ce dernier, nous avons exploré des techniques alternatives permettant, à partir d'informations partielles sur le contenu des documents, de déterminer les concepts pertinents pour les représenter. Pour cela, nous avons proposé deux principales stratégies et une troisième combinant les deux. La première proposition est basée sur l'algorithme des k plus proches voisins et la seconde est basée sur l'analyse sémantique explicite.

Après leur évaluation sur une large collection standard, nous avons constaté que la première proposition permet d'obtenir des performances comparables aux résultats de l'état de l'art actuel, tandis que les résultats de la seconde sont plus faibles. En conséquence, la technique ESA ne semble pas adaptée pour ce type de classification, contrairement à l'algorithme k-NN que nous envisageons d'explorer davantage pour améliorer encore les résultats. Nous avons

aussi testé la combinaison des deux approches mais elle n'a pas permis d'améliorer les résultats.

Pour améliorer la méthode de recherche de voisins, nous comptons explorer d'autres mesures, telles que le coefficient de corrélation de Pearson et le coefficient de Jaccard qui se sont montrés plus efficaces que la mesure du cosinus dans certaines expérimentations (Huang, 2008). Il serait également intéressant d'explorer d'autres modèles de RI tels que le BM25 (Robertson et Walker, 1994) ou encore le modèle de langue (Ponte et Croft, 1998).

Concernant la recherche de documents similaires, nous nous sommes basés sur la mesure du cosinus qui est largement utilisée en RI. Il serait intéressant d'investiguer la combinaison de cette dernière avec des ressources de connaissances du domaine afin d'améliorer la méthode de calcul de similarité basée seulement sur les mots communs.

Chapitre 6: Discussion et perspectives

Dans ce chapitre, nous discutons les limites de notre travail et de quelques-unes de ses perspectives. Nous abordons tout d'abord des questions encore en suspens concernant la construction d'ontologies puis nous examinons des points restant à étudier dans le cadre de la RI sémantique.

1 Construction d'ontologies

Les ontologies sont un outil important pour la représentation des connaissances. Différentes expérimentations en RI ont montré également leur intérêt pour l'amélioration des performances des SRI. Toutefois, leur mise en place est souvent fastidieuse et coûteuse. Dans la première partie de ce travail, nous avons proposé une méthodologie basée sur la réutilisation de RTO existantes pour alléger le processus de construction d'ontologies. Cette dernière soulève un certain nombre de questions que nous exposons ci-après.

1.1 Intérêts de la réutilisation de ressources termino-ontologiques existantes

Dans l'ingénierie ontologique, la réutilisation de ressources sémantiques existantes est un point important qui a été largement exploré. L'approche que nous avons proposée présente un certain nombre d'avantages. En effet, si la plupart des méthodes de construction d'ontologies basées sur cette technique se contentent seulement des connaissances contenues dans la ressource (Hahn et Schulz, 2004), la nôtre, en plus de ces connaissances, exploite des corpus de textes pour les compléter. En plus, l'enrichissement de l'ontologie avec l'intégration de nouvelles entités (termes, concepts, relations) se fait de manière semi-automatique, contrairement à la plupart des méthodes où cette étape est effectuée manuellement (Bontas et al., 2005; Jiménez-Ruiz et al., 2008). En outre, certaines approches considèrent que les sources de connaissances exploitées sont consistantes et, par conséquent, n'abordent pas le traitement d'éventuelles incohérences. Dans notre cas, des règles ont été introduites pour traiter les inconsistances et redondances existant dans les connaissances extraites. D'autres travaux considèrent les relations hiérarchiques originales comme des relations taxonomiques (Chrisment et al., 2008; Jiménez-Ruiz et al., 2008) alors que ces dernières ne sont pas forcément des relations de taxonomie à proprement parler.

Par ailleurs, notre approche se veut applicable dans un contexte multilingue. Nous avons ainsi proposé une méthode d'alignement de termes combinant des techniques heuristique et statistique. Pour ce faire, la méthode proposée dans (Drame et al., 2012) a été combinée à une approche exploitant le traducteur statistique Moses, toutes deux basées sur des corpus parallèles. Notons cependant que l'acquisition de corpus parallèles appropriés n'est pas une tâche triviale, notamment dans des domaines spécifiques comme la maladie d'Alzheimer ; ce qui peut limiter l'applicabilité de notre approche à des domaines où de telles ressources ne sont pas disponibles.

1.2 Généralisabilité de notre approche

Un aspect très important dans l'ingénierie ontologique est la reproductibilité des méthodologies de construction d'ontologies. Elle assure leur applicabilité à des domaines différents. Dans l'approche que nous proposons, les différents outils et techniques d'extraction d'information utilisés sont indépendants d'un domaine spécifique. Cette étape peut ainsi être exploitée dans n'importe quel domaine à condition de disposer d'un corpus de textes suffisamment représentatif. La structuration des concepts est une phase dépendant de l'UMLS et comme ce système couvre largement les connaissances biomédicales, elle peut être appliquée pour la construction et/ou l'enrichissement d'ontologies dans n'importe quel sous-domaine biomédical. Pour être transposée à d'autres domaines, il est nécessaire que des RTO décrivant leurs connaissances soient disponibles, comme souligné dans (Bontas et al., 2005). En ce qui concerne la phase d'alignement de termes exprimés dans des langues différentes, notre approche est complètement indépendante du domaine et est donc totalement généralisable, moyennant l'existence de corpus parallèles ou de sites Internet multilingues présentant le même contenu. L'enrichissement de l'ontologie étant basé sur les dépendances syntaxiques fournies par Syntex, l'intégration de nouveaux concepts est donc possible quel que soit le domaine d'intérêt.

1.3 Structuration des concepts de l'ontologie

Pour structurer les concepts, nous avons, dans un premier temps, extrait des relations taxonomiques définies explicitement dans l'UMLS. Nous avons ensuite exploité les relations hiérarchiques plus vagues (CHILD / PARENT / BROADER / NARROWER) pour tenter de relier les concepts isolés aux autres concepts de l'ontologie.

Concernant l'intégration des nouveaux concepts, les dépendances en tête entre les termes les dénotant et les entrées de l'ontologie ont été exploitées pour les intégrer dans la hiérarchie. Cependant, les nouveaux concepts pour lesquels il n'a pas été possible d'établir des dépendances syntaxiques, ont été rattachés directement à la racine de l'ontologie. Ainsi, de trop nombreux concepts se trouvent au premier niveau de celle-ci. Concernant les concepts retrouvés dans l'UMLS, leurs types sémantiques sont utilisés pour mieux les structurer tandis que pour les nouveaux, une telle information n'est pas disponible. Ainsi, il serait intéressant de mettre en place des techniques additionnelles permettant de positionner correctement les nouveaux concepts au sein de l'ontologie. Dans (Chrisment et al., 2008), les auteurs proposent la définition de concepts génériques (types abstraits) afin de regrouper ces concepts et de mieux les structurer. Nous envisageons de tester leur approche sur les types sémantiques de l'UMLS. Par ailleurs, nous souhaiterions explorer l'utilisation de patrons morphosyntaxiques mais également les approches distributionnelles qui, en fonction des contextes d'utilisation, permettent de lier les concepts.

D'autre part, nous conservons les relations associatives définies dans les vocabulaires sources de l'UMLS. Ces dernières, bien qu'étant typées, sont généralement vagues. Par exemple, la relation *clinically_associated_with* lie plus de 1600 couples de concepts et associe le concept *Dementia* à 22 autres concepts (*Cerebrovascular accident*, *Hypertensive disease*, *Vitamin B*

12, Neuraxis), La sémantique d'une telle relation est variable. Elle nécessiterait une spécification détaillée.

1.4 Vers une ontologie multilingue de la maladie d'Alzheimer

Une importante question abordée dans notre travail est l'aspect multilingue. Dans le but de mettre en place une ontologie bilingue (français-anglais), une méthode d'alignement de termes basée sur les corpus parallèles a été développée. Cette dernière a permis d'enrichir les aspects multilingues et terminologiques de l'ontologie avec la découverte de nouveaux synonymes dans les deux langues (français-anglais). Pour permettre à la communauté scientifique internationale de partager facilement et largement les connaissances sur la maladie d'Alzheimer, l'ontologie pourrait être étendue pour supporter d'autres langues. Puisque la ressource externe que nous avons exploitée (l'UMLS) est multilingue, elle pourrait être utilisée à cette fin. De plus, notre méthode d'alignement n'étant pas spécifique à des couples de langues particulières, elle permet d'aligner des termes à partir de corpus parallèles pour n'importe quelle paire de langues. Elle peut ainsi servir pour l'extension de l'ontologie à d'autres langues. Toutefois, sa limite reste l'acquisition de corpus parallèles adaptés sur lesquels elle se base. Une piste intéressante est l'exploration des techniques basées sur les corpus comparables qui sont de plus en plus utilisés pour la construction de ressources multilingues (Morin et Prochasson, 2011; Hazem et Morin, 2012; Bouamor et al., 2013). En particulier, les analyses critiques réalisées par les experts de BiblioDem sur les articles scientifiques concernant la maladie d'Alzheimer pourraient être associés aux résumés correspondants pour constituer un corpus comparable.

1.5 Evaluation de l'ontologie

L'évaluation d'ontologies est toujours une question de recherche ouverte. Dans le cadre de ce travail, une validation de l'ontologie dans un cadre pratique est en cours. Différents aspects seront évalués : la pertinence de l'ontologie pour l'application cible, sa couverture pour le domaine, etc. En effet, l'ontologie résultante a été raffinée et validée par des experts du domaine mais elle n'a pas encore été évaluée dans une application pratique ou par d'autres spécialistes de la maladie d'Alzheimer. Elle supporte actuellement le portail SemBiP et une évaluation orientée utilisation, dont nous avons présenté des résultats préliminaires, est en cours. L'objectif est de trouver des réponses aux questions suivantes :

- L'ontologie permet-elle aux utilisateurs du portail d'accéder facilement et rapidement aux documents correspondant à leurs besoins en information ?
- L'ontologie couvre-t-elle suffisamment les connaissances du domaine ?
  Permet-elle de prendre en compte la variété des utilisateurs et leurs niveaux d'expertise ?

Nous envisageons aussi, pour améliorer le degré de formalisation de notre ontologie et faciliter son interopérabilité avec d'autres ontologies biomédicales, de l'aligner avec l'ontologie de haut niveau BFO (Grenon et al., 2004). Celle-ci est basée sur la distinction entre les « entités qui perdurent au fil du temps » et les « entités qui se produisent pour une durée ». En s'appuyant sur ce principe, elle définit des concepts de haut niveau, indépendants

d'un domaine spécifique (*object*, *function*, *process*) qui peuvent être liés aux concepts se trouvant au premier niveau de notre ontologie.

2 Recherche d'information sémantique

La RI sémantique s'est montrée prometteuse et a fait l'objet de nombreuses investigations. L'intérêt suscité par les premiers travaux dans le domaine a motivé le développement de cette approche dans le domaine biomédical où les ressources sont multiples (Zhou et al., 2006b). Vu la variabilité de la terminologie biomédicale, la forte présence d'acronymes et d'abréviations dans le domaine, mais aussi la diversité des données (Zweigenbaum et al., 2001), la RI sémantique a soulevé de nombreux challenges, tels que l'identification des entités médicales dans des textes, leur désambiguïsation ou encore l'incomplétude des sources de connaissances. Dans ce travail, nous nous sommes intéressés à ces différentes questions et avons proposé des méthodes pour les traiter. Ce travail présente toutefois quelques limites. Nous exposons ces limites et présentons certaines pistes pour les surmonter.

2.1 Limites de la méthode d'indexation

Concernant l'extraction des concepts, la méthode utilisant le chunking reste limitée pour le traitement de documents cliniques car elle ignore beaucoup d'entités médicales, qui n'ont pas la structure d'un syntagme nominal. Notre approche basée sur les n-grammes permet de surmonter cette limite mais soulève d'autres problèmes. En considérant tout n-gramme comme un candidat terme, cette dernière génère beaucoup de bruit. Par exemple, pour l'expression *démence et corps de Lévy*, elle va identifier toute la chaine de caractères comme un seul concept car, après normalisation, cette dernière est alignée avec le concept *démence à corps de Lévy*. Cette méthode est confrontée aux problèmes liés aux mots de liaisons (e.g., et, ou, avec) comme soulignés aussi dans l'extraction de syntagmes nominaux (Kang et al., 2011). Des méthodes statistiques basées sur l'apprentissage automatique, telles que le CRF, pourraient être une solution à ce type de problème. Mais pour la mise en œuvre de telles méthodes, des données annotées au préalable sont requises.

Pour la désambiguïsation des termes, nous avons proposé un algorithme basé sur la similarité sémantique entre les concepts. L'évaluation montre que l'intégration de la désambiguïsation dans la phase d'identification de concepts améliore significativement les résultats. Mais cette méthode n'a pas été testée sur une tâche spécifique de désambiguïsation. Il serait ainsi intéressant de l'évaluer et de la comparer aux méthodes de l'état de l'art pour voir l'impact de la prise en compte du niveau d'ambiguïté des termes du contexte d'un terme ambigu.

2.2 Vers un système de recherche d'information « intelligent »

Les ontologies, bien que permettant d'améliorer les performances en RI, restent sous-exploitées dans notre approche. Malgré le potentiel des ontologies, notre modèle exploite seulement des concepts qu'elles englobent et des liens hiérarchiques entre ces derniers. Dans la littérature, des travaux ont proposé, en plus des ontologies, l'exploitation de bases de connaissances sous-jacentes dans la RI (Kiryakov et al., 2004; Fernández et al., 2011). L'utilisation de bases de connaissances pourraient permettre de disposer de systèmes de RI

plus « intelligents ». Par exemple, pour la requête « *quels sont les facteurs de risque de la maladie d'Alzheimer identifiés chez les personnes habitant dans le Sud-Ouest de la France* », le système, au lieu de retourner les documents contenant les concepts de la requête ou des concepts similaires et approchants, pourrait lui proposer les facteurs de risque de cette maladie identifiés dans la collection de documents. Le développement de tels systèmes requiert toutefois la mise en place de ces bases, mais aussi de systèmes d'annotations permettant d'identifier les assertions appropriées dans des documents. Les systèmes de Questions/Réponses permettent de répondre à ce type de questions et constituent une perspective intéressante.

2.3 Vers un modèle de recherche d'information personnalisé

Le portail SemBiP est destiné à une variété d'utilisateurs ayant des profils et des niveaux d'expertise différents (décideurs politiques, praticiens, chercheurs, étudiants, utilisateurs grand public, etc.). Ils peuvent ainsi avoir des besoins en information différents ; avec une même requête, ils peuvent avoir des attentes différentes. La prise en compte du profil de l'utilisateur est une question importante dans notre contexte qui mériterait d'être prise en compte. Un modèle de RI personnalisé permettrait à ces utilisateurs de réaliser des recherches pertinentes et adaptées à leurs connaissances du domaine et donc d'améliorer les résultats du système. Une option simple serait que les utilisateurs puissent lors de leur première utilisation du portail, indiquer leurs champs de compétence ou domaines d'intérêt. Les concepts de l'ontologie pouvant être associés aux informations fournies par les utilisateurs permettraient, par la suite, de définir leurs profils et ces derniers pourront aussi être complétés et adaptés au fur et à mesure de leur navigation. Enfin, toutes ces informations, qui peuvent être stockées dans une base de données ou une ontologie, permettront de filtrer ou d'étendre les résultats de la recherche en fonction du profil de l'utilisateur.

2.4 Application de KNN-Classifier pour l'attribution d'articles aux relecteurs de BiblioDémences

Dans le chapitre 5, nous avons proposé une méthode de classification de documents basée sur l'approche des k-NN. Cette dernière permet, à partir d'un ensemble de documents préalablement annotés, de déterminer les concepts pertinents pour annoter un document donné. Elle pourrait ainsi servir pour l'attribution des nouveaux documents aux relecteurs qui doivent ensuite en faire une analyse critique. Actuellement, cette tâche est faite manuellement en fonction de l'expertise des relecteurs. Tous les articles intégrés dans la base BiblioDem sont associés aux spécialistes qui ont réalisé leurs analyses. Cet ensemble pourrait servir comme une collection d'entrainement aux classifieurs et sur laquelle se baser pour l'attribution (semi-)automatique de nouveaux articles aux relecteurs en fonction de leur profil.

Conclusions

L'objectif de ce travail était d'explorer le potentiel des ontologies pour améliorer les performances des modèles de RI. Ainsi, nous nous sommes intéressés à deux thématiques : la construction d'ontologies et leur utilisation dans la RI.

Dans un premier temps, nous avons effectué une revue des approches de construction d'ontologies représentatives. Ensuite, en nous inspirant de ces travaux, nous avons proposé une approche de construction d'ontologies combinant l'extraction d'informations à partir de textes et la réutilisation de RTO existantes. Ces dernières étant particulièrement nombreuses dans le domaine biomédical, leur utilisation devait en principe permettre d'alléger le processus de construction d'ontologies. D'abord, un outil de TAL a été utilisé pour extraire les termes candidats et les relations syntaxiques qu'ils entretiennent. Ensuite, la vaste ressource biomédicale UMLS a permis de regrouper ces termes en concepts et de structurer ces derniers. Cette ressource présentant des incohérences, nous avons réalisé des traitements supplémentaires, d'une manière semi-automatique, afin d'améliorer la représentation des connaissances extraites. L'approche de construction d'ontologies proposée intègre également une méthode d'alignement permettant d'apparier des termes dans des langues différentes. Elle est donc adaptée pour l'application dans des contextes multilingues. Cette méthodologie est dédiée au développement d'ontologies dites légères, qui sont suffisantes pour les tâches de RI (Chrisment et al., 2008). Elle a été exploitée pour mettre en œuvre une ontologie bilingue de la maladie d'Alzheimer, OntoAD, et est actuellement utilisée pour supporter un moteur de recherche sémantique bilingue implémenté au sein du portail SemBiP.

Notre deuxième contribution se situe dans le champ de la RI sémantique, utilisant une ontologie. Nous avons passé en revue les travaux menés dans ce domaine et avons montré qu'ils ont permis de surmonter les limites des méthodes classiques de RI, basées sur les mots clés, et en particulier les problèmes liés à l'ambiguïté et à la disparité des termes. L'utilisation des concepts plutôt que des termes permet aussi de se départir de la barrière de la langue, en rendant possible la gestion de plusieurs langues grâce aux termes synonymes qui peuvent être associés aux concepts. Les liens sémantiques entre les concepts peuvent également être exploités pour améliorer l'appariement entre les documents et les requêtes. De plus, les ressources sémantiques sur lesquelles les modèles de RI s'appuient sont largement disponibles dans le domaine biomédical.

Notre contribution dans ce domaine se situe à deux niveaux. Tout d'abord, nous avons proposé deux approches de repérage de concepts en vue de l'indexation conceptuelle de documents et une méthode de désambiguïsation basée sur la similarité sémantique. Ces deux approches, l'une basée sur les syntagmes nominaux et l'autre sur les n-grammes ont été évaluées sur deux corpus standard et ont donné des résultats prometteurs. La méthode basée sur les n-grammes étant plus performante, nous l'avons ainsi utilisée pour mettre en œuvre le portail sémantique SemBiP. En plus d'une fonctionnalité de recherche conceptuelle, SemBiP supporte la combinaison de la recherche sémantique et de la recherche par mots clés mais également l'expansion de requêtes. Une évaluation orientée utilisateur du portail montre que

les utilisateurs y portent un grand intérêt. En plus, leurs retours sont globalement positifs avec des suggestions pour rendre le portail plus fonctionnel.

Ensuite, nous avons proposé, étant donné une ressource sémantique, deux stratégies pour la classification multi-labels d'un large corpus de documents biomédicaux, et ce, afin de faciliter la tâche d'annotation. La première stratégie est basée sur l'algorithme des k plus proches voisins dont le principe est, pour déterminer les concepts permettant d'annoter un document donné, de considérer ceux associés aux documents qui lui sont similaires. Pour cela, nous avons exploré différentes configurations dont celle combinant le modèle vectoriel pour la recherche de documents proches, la méthode Random Forest pour la classification des concepts et la technique de comparaison de scores de concepts de rangs successifs pour fixer le nombre de labels. La seconde stratégie utilise l'analyse sémantique explicite en adaptant la manière dont les associations entre mots et concepts sont capturées. Pour cela, nous avons utilisé la mesure de Jaccard afin de capturer ces associations à partir d'une large collection de documents. Ainsi, pour chaque concept, les termes qui lui sont le plus fortement associés sont déterminés et utilisés pour l'indexer. Ensuite, pour tout document à annoter (requête), on retrouve par ordre décroissant de correspondance les concepts (documents) les plus pertinents en utilisant le modèle vectoriel. Ces deux stratégies ont été évaluées sur des collections de documents biomédicaux fournies lors du challenge BioASQ dans son édition 2014, où les documents étaient annotés avec les descripteurs du thésaurus MeSH (Medical Subject Heading). Nous avons montré que la première proposition obtenait de bonnes performances avec des résultats comparables (et parfois meilleurs que) aux meilleurs systèmes actuellement décrits dans la littérature tandis que la deuxième proposition, bien qu'étant capable de traiter une collection importante de documents (plus de quatre millions de documents) dans un temps raisonnable (deux heures) s'est avérée moins performante.

Publications

Revues internationales

K. Dramé, G. Diallo, F. Delva, J. F. Dartigues, E. Mouillet, R. Salamon, F. Mougin. ***Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application to Alzheimer's disease***. Journal of Biomedical Informatics, Volume 48, April 2014, Pages 171-182, http://dx.doi.org/10.1016/j.jbi.2013.12.013.

K. Dramé, F. Mougin, F. Delva, J. F. Dartigues, E. Mouillet, R. Salamon, G. Diallo. ***Large scale biomedical documents classification : a k-NN-based approach*** (en préparation pour *Journal of Biomedical Semantics*).

Conférences et workshops internationaux

K. Dramé, G. Diallo, F. Mougin. ***Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus***. 24th European Medical Informatics Conference - MIE2012. 26-29th August 2012, Pisa, Italy.

K. Dramé, F. Mougin, G. Diallo. ***Query expansion using external resources for improving information retrieval in the biomedical domain***, CLEF/eHealth 2014.

K. Dramé, F. Mougin, G. Diallo. ***A k-nearest neighbor based method for improving large scale biomedical document indexing***, 6th International Symposium on Semantic Mining in Biomedicine (SMBM), 6th-7th October, 2014, University of Aveiro, Portugal.

Conférences et workshops nationaux

K. Dramé, G. Diallo, F. Mougin. ***Réutilisation de ressources terminologiques existantes pour la construction d'ontologie de domaine*.** CNRIA 2013, 24-27 Avril 2013 à Ziguinchor, Sénégal.

K. Dramé, G. Diallo, F. Mougin. ***Construction d'une ontologie bilingue de la maladie d'Alzheimer à partir de textes médicaux***. Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé ; 26 Juin 2012, Paris, France.

K. Dramé, G. Diallo, F. Mougin. ***Conception et utilisation d'ontologies pour l'accès aux informations pertinentes (poster)***. EARIA 2012, 24-26 octobre 2012 à la Tourette, France, Poster.

G. Diallo, K. Dramé, F. Delva, JF. Dartigues, F. Mougin. ***Representing and accessing scientific knowledge about the Alzheimer's Disease: the Semantic BiblioDem Portal***.4th ICST International Conference on eHealth (eHealth 2011), 21-23 Nov 2011, Malaga, Spain, Poster.

Bibliographie

Abacha, A.B., Zweigenbaum, P., 2011. Medical Entity Recognition: A Comparison of Semantic and Statistical Methods, in: Proceedings of BioNLP 2011 Workshop, BioNLP '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 56–64.

Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-Based Learning Algorithms. Mach Learn 6, 37–66. doi:10.1023/A:1022689900470

Alani, H., 2006. Position Paper: Ontology Construction from Online Ontologies, in: Proceedings of the 15th International Conference on World Wide Web, WWW '06. ACM, New York, NY, USA, pp. 491–495. doi:10.1145/1135777.1135849

Aronson, A., Lang, F.-M., 2010. An overview of MetaMap: historical perspective and recent advances. J. Am. Med. Inform. Assoc. JAMIA 17, 229–236. doi:10.1136/jamia.2009.002733

Aronson, A.R., 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program.

Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J., 2004. The NLM Indexing Initiative's Medical Text Indexer. Stud. Health Technol. Inform. 107, 268–272.

Aronson, A.R., Rindflesch, T.C., 1997. Query expansion using the UMLS Metathesaurus. Proc. Conf. Am. Med. Inform. Assoc. AMIA Annu. Fall Symp. AMIA Fall Symp. 485–489.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Assem, M.V., Menken, M.R., Schreiber, G., Wielemaker, J., Wielinga, B., 2004. A Method for Converting Thesauri to RDF/OWL, in: Proc. of the 3rd Int'l Semantic Web Conf. (ISWC'04), Number 3298 in Lecture Notes in Computer Science. Springer-Verlag, pp. 17–31.

Aubin, S., Hamon, T., 2006. Improving Term Extraction with Terminological Resources, in: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (Eds.), Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August 23-25, 2006, Proceedings, Lecture Notes in Computer Science. Springer, pp. 380–387. doi:10.1007/11816508_39

Aussenac-Gilles, N., Condamines, A., Szulman, S., 2002. Prise en compte de l'application dans la constitution de produits terminologiques, in: 2e Assises Nationales Du GDR I3, Nancy (F), 04/12/2002-06/12/2002. Cepadeus Editions, Toulouse (F), pp. 289–302.

Aussenac-Gilles, N., Despres, S., Szulman, S., 2008. The TERMINAE Method and Platform for Ontology Engineering from Texts, in: Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge. IOS Press, Amsterdam, The Netherlands, The Netherlands, pp. 199–223.

Azcárate, M.C., Vázquez, J.M., López, M.M., 2012. Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. J. Am. Med. Inform. Assoc. amiajnl–2012–000943. doi:10.1136/amiajnl-2012-000943

Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (Eds.), 2003. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York, NY, USA.

Baeza-Yates, R.A., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Balikas, G., Partalas, I., Ngomo, A.-C.N., Krithara, A., Paliouras, G., 2014. Results of the BioASQ Track of the Question Answering Lab at CLEF 2014, in: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings. CEUR-WS.org, pp. 1181–1193.

Batet, M., Sánchez, D., Valls, A., 2011. An Ontology-based Measure to Compute Semantic Similarity in Biomedicine. J Biomed. Inform. 44, 118–125. doi:10.1016/j.jbi.2010.09.002

Baziz, M., Boughanem, M., Aussenac-Gilles, N., 2005. Conceptual Indexing Based on Document Content Representation, in: Crestani, F., Ruthven, I. (Eds.), Context: Nature, Impact, and Role, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 171–186.

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web: Scientific American. Sci. Am.

Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D., 2008. Hybrid search: effectively combining keywords and semantic searches, in: Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications, ESWC'08. Springer-Verlag, Berlin, Heidelberg, pp. 554–568.

Bhogal, J., Macfarlane, A., Smith, P., 2007. A review of ontology based query expansion. Inf. Process. Manag. 43, 866–886. doi:10.1016/j.ipm.2006.09.003

Biébow, B., Szulman, S., 1999. TERMINAE: A Linguistics-Based Tool for the Building of a Domain Ontology. Knowl. Acquis. Model. Manag. 11th Eur. Workshop EKAW 99 Dagstuhl Castle Ger. May 26-29 1999 Proc. 1621, 49–66.

Bodenreider, O., 2001. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention.

Bodenreider, O., 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, 267–270.

Bodenreider, O., 2006. Lexical, terminological and ontological resources for biological text mining, in: Ananiadou, S., McNaught, J. (Eds.), Text Mining for Biology and Biomedicine. Artech House, pp. 43–66.

Bodenreider, O., Bodenreider, and O., Mccray, A.T., 2003. Exploring Semantic Groups Through Visual Approaches. J. Biomed. Inform. 36, 414–432.

Bodenreider, O., Mitchell, J.A., McCray, A.T., 2002. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc. AMIA Symp. 61–65.

Bontas, E.P., Mochol, M., Tolksdorf, R., 2005. Case studies on ontology reuse, in: In Proceedings of the 5th International Conference on Knowledge Management IKNOW05.

Bouamor, D., Semmar, N., Zweigenbaum, P., 2013. Context Vector Disambiguation for Bilingual Lexicon Extraction from Comparable Corpora, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers. The Association for Computer Linguistics, pp. 759–764.

Bourigault, D., 1994. Lexter : un Logiciel d'EXtraction de TERminologie : application à l'acquisition des connaissances à partir de textes. EHESS.

Bourigault, D., Aussenac-Gilles, N., 2003. Construction d'ontologies à partir de textes, in: TALN , Batz-Sur-Mer.

Bourigault, D., Fabre, C., 2000. Approche linguistique pour l'analyse syntaxique de corpus. Cah. Gramm. 25, 131–151.

Breiman, L., 2001. Random Forests. Mach Learn 45, 5–32. doi:10.1023/A:1010933404324

Buckley, C., Singhal, A., Mitra, M., 1995. New Retrieval Approaches Using SMART: TREC 4. Presented at the Proceedings of the Fourth Text Retrieval Conference (TREC-4), pp. 25–48.

Buitelaar, P., Magnini, B., 2005. Ontology Learning from Text: An Overview, in: In Paul Buitelaar, P., Cimiano, P., Magnini B. (Eds.), Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press, pp. 3–12.

Buitelaar, P., Olejnik, D., Sintek, M., 2004. A Protege Plug-in for Ontology Extraction from Text Based on Linguistic Analysis, in: In Proceedings of the 1st European Semantic Web Symposium (ESWS.

Camara, G., Desprès, S., Djedidi, R., Lo, M., 2013. Design of Schistosomiasis Ontology (IDOSCHISTO) Extending the Infectious Disease Ontology, in: Lehmann, C.U., Ammenwerth, E., Nøhr, C. (Eds.), MEDINFO 2013 - Proceedings of the 14th World Congress on Medical and Health Informatics, 20-13 August 2013, Copenhagen, Denmark, Studies in Health Technology and Informatics. IOS Press, pp. 466–470. doi:10.3233/978-1-61499-289-9-466

Campos, D., Matos, S., Oliveira, J.L., 2013. Gimli: open source and high-performance biomedical name recognition. BMC Bioinformatics 14, 54. doi:10.1186/1471-2105-14-54

Castells, P., Fernandez, M., Vallet, D., 2007. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Trans. Knowl. Data Eng. 19, 261–272. doi:10.1109/TKDE.2007.22

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., Vandenbussche, P.-Y., others, 2012. Construire une ontologie médicale pour la recherche d'information: problématiques terminologiques et de modélisation. Actes 23es Journ. Francoph. Ingénierie Connaiss. 33–48.

Chrisment, C., Haemmerlé, O., Hernandez, N., Mothe, J., 2008. Méthodologie de transformation d'un thesaurus en une ontologie de domaine. Rev. Intell. Artif. 22, 7–37.

Ciccarese, P., Wu, E., Wong, G., Ocana, M., Kinoshita, J., Ruttenberg, A., Clark, T., 2008. The SWAN biomedical discourse ontology. J. Biomed. Inform. 41, 739–751.

Cimiano, P., Völker, J., 2005. Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery.

Cimino, J.J., Min, H., Perl, Y., 2003. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J. Biomed. Inform. 36, 450–461. doi:10.1016/j.jbi.2003.11.001

Contreras, J., Benjamins, V.R., Blázquez, M., Losada, S., Salla, R., Sevilla, J., Navarro, D., Casillas, J., Mompó, A., Patón, D., Corcho, O., Tena, P., Martos, I., 2004. A Semantic Portal for the International Affairs Sector, in: Motta, E., Shadbolt, N.R., Stutt, A., Gibbins, N. (Eds.), Engineering Knowledge in the Age of the Semantic Web, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 203–215.

Croft, W.B., 1986. User-specified Domain Knowledge for Document Retrieval, in: Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '86. ACM, New York, NY, USA, pp. 201–206. doi:10.1145/253168.253211

Daille, B., 1994. Approche mixte pour l''extraction de terminologie : statistique lexicale et filtres linguistiques.

D'Aquin, M., Lewen, H., 2009. Cupboard — A Place to Expose Your Ontologies to Applications and the Community, in: Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications, ESWC 2009 Heraklion. Springer-Verlag, Berlin, Heidelberg, pp. 913–918. doi:10.1007/978-3-642-02121-3_81

D'Aquin, M., Motta, E., 2011. Watson, More Than a Semantic Web Search Engine. Semantic Web 2, 55–63.

D'Aquin, M., Noy, N.F., 2012. Where to publish and find ontologies? A survey of ontology libraries. Web Semant. Sci. Serv. Agents World Wide Web 11, 96–111. doi:10.1016/j.websem.2011.08.005

D'Aquin, M., Sabou, M., Motta, E., 2008. Reusing Knowledge from the Semantic Web with the Watson Plugin, in: Bizer, C., Joshi, A. (Eds.), Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008, CEUR Workshop Proceedings. CEUR-WS.org.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990a. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990b. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41, 391–407.

Déjean, H., Gaussier, E., Renders, J.-M., Sadat, F., 2005. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. Artif Intell Med 33, 111–124. doi:10.1016/j.artmed.2004.07.015

De Keizer, N.F., Abu-Hanna, A., Zwetsloot-Schonk, J.H., 2000. Understanding terminological systems. I: Terminology and typology. Methods Inf. Med. 39, 16–21.

Deléger, L., Grouin, C., Zweigenbaum, P., 2010. Extracting medical information from narrative patient records: the case of medication-related information. JAMIA 17, 555–558. doi:10.1136/jamia.2010.003962

Deléger, L., Merkel, M., Zweigenbaum, P., 2009. Translating medical terminologies through word alignment in parallel text corpora. J Biomed. Inform. 42, 692–701. doi:10.1016/j.jbi.2009.03.002

Diallo, G., 2006. Une architecture à base d'ontologies pour la gestion unifiée des données structurées et non structurées. Thèse : Université Joseph Fourier – Grenoble I**.**

Diallo, G., 2011. Efficient Building of Local Repository of Distributed Ontologies, in: 2011 Seventh International Conference on Signal-Image Technology and Internet-Based Systems (SITIS). Presented at the 2011 Seventh International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), pp. 159–166. doi:10.1109/SITIS.2011.45

Díaz-Galiano, M.C., Martín-Valdivia, M.T., Ureña-López, L.A., 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. Comput. Biol. Med. 39, 396–403. doi:10.1016/j.compbiomed.2009.01.012

Diem, L.T.H., Chevallet, J.-P., Thuy, D.T.B., 2007. Thesaurus-based query and document expansion in conceptual indexing with UMLS: Application in medical information retrieval, in: 2007 IEEE International Conference on Research, Innovation and Vision for the Future. Presented at the 2007 IEEE International Conference on Research, Innovation and Vision for the Future, pp. 242–246. doi:10.1109/RIVF.2007.369163

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R.S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J., 2004. Swoogle: A Search and Metadata Engine for the Semantic Web, in: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge

Management, CIKM '04. ACM, New York, NY, USA, pp. 652–659. doi:10.1145/1031171.1031289
Ding, Y., Fensel, D., 2001. Ontology Library Systems: The key to successful Ontology Re-use, in: Stanford University 2001; S. pp. 93–112.
Ding, Y., Sun, Y., Chen, B., Borner, K., Ding, L., Wild, D., Wu, M., DiFranzo, D., Fuenzalida, A.G., Li, D., Milojevic, S., Chen, S., Sankaranarayanan, M., Toma, I., 2010. Semantic Web Portal: A Platform for Better Browsing and Visualizing Semantic Data, in: An, A., Lingras, P., Petty, S., Huang, R. (Eds.), Active Media Technology, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 448–460.
Dinh, B.-D., 2012. Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques (phd). Université de Toulouse, Université Toulouse III - Paul Sabatier.
Dinh, B., Tamine, L., 2010. Sense-Based Biomedical Indexing and Retrieval. Presented at the NLDB, pp. 24–35.
Doms, A., Schroeder, M., 2005. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res. 33, W783–W786. doi:10.1093/nar/gki470
Drame, K., Diallo, G., Mougin, F., 2012. Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus, in: Mantas, J., Andersen, S.K. er, Mazzoleni, M.C., Blobel, B., Quaglini, S., Moen, A. (Eds.), Quality of Life through Quality of Information - Proceedings of MIE2012, The XXIVth International Congress of the European Federation for Medical Informatics, Pisa, Italy, August 26-29, 2012, Studies in Health Technology and Informatics. IOS Press, pp. 179–183. doi:10.3233/978-1-61499-101-4-179
Dramé, K., Mougin, F., Diallo, G., 2014. Query Expansion using External Resources for Improving Information Retrieval in the Biomedical Domain, in: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings. CEUR-WS.org, pp. 189–194.
Dumais, S.T., 1994. Latent Semantic Indexing (LSI) and TREC-2, in: The Second Text REtrieval Conference (TREC-2. pp. 105–115.
Egozi, O., Markovitch, S., Gabrilovich, E., 2011. Concept-Based Information Retrieval Using Explicit Semantic Analysis. ACM Trans Inf Syst 29, 8:1–8:34. doi:10.1145/1961209.1961211
Farquhar, A., Fikes, R., Rice, J., 1997. The Ontolingua Server: a tool for collaborative ontology construction. Int J Hum-Comput Stud 46, 707–727.
Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. MIT Press.
Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., Motta, E., 2011. Semantically Enhanced Information Retrieval: An Ontology-based Approach. Web Semant 9, 434–452. doi:10.1016/j.websem.2010.11.003
Fortuna, B., Grobelnik, M., Mladenic, D., 2007. OntoGen: Semi-automatic Ontology Editor, in: Proceedings of the 2007 Conference on Human Interface: Part II. Springer-Verlag, Berlin, Heidelberg, pp. 309–318.
Fortuna, B., Mladenič, D., Grobelnik, M., 2006. Semi-automatic Construction of Topic Ontologies, in: Ackermann, M., Berendt, B., Grobelnik, M., Hotho, A., Mladenič, D., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., Someren, M. van (Eds.), Semantics, Web and Mining, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 121–131.
Frantzi, K.T., Ananiadou, S., Tsujii, J., 1998. The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms, in: Nikolaou, C., Stephanidis, C. (Eds.), Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL

’98, Heraklion, Crete, Greece, September 21-23, 1998, Proceedings, Lecture Notes in Computer Science. Springer, pp. 585–604. doi:10.1007/3-540-49653-X_35
Gabrilovich, E., Markovitch, S., 2006. Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, in: Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2, AAAI’06. AAAI Press, Boston, Massachusetts, pp. 1301–1306.
Gabrilovich, E., Markovitch, S., 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in: In Proceedings of the 20th International Joint Conference on Artificial Intelligence. pp. 1606–1611.
Gale, W.A., Church, K.W., 1991. A program for aligning sentences in bilingual corpora, in: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL ’91. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 177–184. doi:10.3115/981344.981367
Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L., 2002. Sweetening Ontologies with DOLCE, in: Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, EKAW ’02. Springer-Verlag, London, UK, UK, pp. 166–181.
Garla, V., Brandt, C., 2012. Knowledge-Based Biomedical Word Sense Disambiguation: An Evaluation and Application to Clinical Document Classification, in: 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, HISB 2012, La Jolla, CA, USA, September 27-28, 2012. IEEE Computer Society, p. 22. doi:10.1109/HISB.2012.12
Garla, V., Brandt, C., 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics 13, 261. doi:10.1186/1471-2105-13-261
Getman, A.P., Karasiuk, V.V., 2014. A crowdsourcing approach to building a legal ontology from text. Artif. Intell. Law 22, 313–335. doi:10.1007/s10506-014-9159-1
Giannangelo, K., Fenton, S.H., 2008. SNOMED CT Survey: An Assessment of Implementation in EMR/EHR Applications. Perspect. Health Inf. Manag. AHIMA Am. Health Inf. Manag. Assoc. 5.
Gobeill, J., Ruch, P., Zhou, X., 2009. Query and Document Expansion with Medical Subject Headings Terms at Medical Imageclef 2008, in: Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access, CLEF’08. Springer-Verlag, Berlin, Heidelberg, pp. 736–743.
Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., Mueller, H., 2014. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval, in: Proceedings of CLEF 2014.
Gómez-Pérez, A., 1999. Ontological Engineering: A State Of The Art.
Gonzalo, J., Verdejo, F., Chugur, I., Cigarrin, J., 1998. Indexing with WordNet synsets can improve text retrieval. pp. 38–44.
Grenon, P., Smith, B., 2004. SNAP and SPAN: Towards Dynamic Spatial Ontology. Spat. Cogn. Amp Comput. 4, 69–104. doi:10.1207/s15427633scc0401_5
Grenon, P., Smith, B., Goldberg, L., 2004. Biodynamic Ontology: Applying BFO in the Biomedical Domain, in: Stud. Health Technol. Inform. IOS Press, pp. 20–38.
Gruber, T.R., 1993a. Toward Principles for the Design of Ontologies Used for Knowledge Sharing, in: International Journal of Human-Computer Studies - Special issue: the role of formal ontology in the information technology, Volume 43 Issue 5-6, PP. 907 - 928.
Gruber, T.R., 1993b. A Translation Approach to Portable Ontology Specifications. Knowl Acquis 5, 199–220. doi:10.1006/knac.1993.1008

Gruninger, M., Fox, M.S., 1996. The Role of Competency Questions in Enterprise Engineering.

Guarino, N., 1998. Formal Ontology and Information Systems. IOS Press, pp. 3–15.

Guarino, N., Giaretta, P., 1995. Ontologies and knowledge bases: Towards a terminological clarification, in: Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. IOS Press, pp. 25–32.

Guarino, N., Masolo, C., Vetere, G., 1999. OntoSeek: Content-Based Access to the Web. IEEE Intell. Syst. 14, 70–80.

Gu, H. (Helen), Perl, Y., Elhanan, G., Min, H., Zhang, L., Peng, Y., 2004. Auditing concept categorizations in the UMLS. Artif. Intell. Med. 31, 29–44. doi:10.1016/j.artmed.2004.02.002

Hahn, U., Schulz, S., 2004. Building a very large ontology from medical thesauri. Handb. Ontol. 133–150.

Haldar, R., Mukhopadhyay, D., 2011. Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach. ArXiv11011232 Cs Math.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl 11, 10–18. doi:10.1145/1656274.1656278

Hazem, A., Morin, E., 2012. Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora, in: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012. European Language Resources Association (ELRA), pp. 288–292.

Hearst, M.A., 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, in: Proceedings of the 14th Conference on Computational Linguistics - Volume 2, COLING '92. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 539–545. doi:10.3115/992133.992154

Hepp, M., de Bruijn, J., 2007. GenTax: A generic methodology for deriving OWL and RDF-S ontologies from hierarchical classifications, thesauri, and inconsistent taxonomies. Semantic Web Res. Appl. 129–144.

Hersh, W., Price, S., Donohoe, L., 2000. Assessing thesaurus-based query expansion using the UMLS metathesaurus, in: In Proc. of the 2000 American Medical Informatics Association (AMIA) Symposium. pp. 344–348.

Hersh, W.R., Cohen, A.M., Ruslen, L., Roberts, P.M., 2007. TREC 2007 Genomics Track Overview, in: Voorhees, E.M., Buckland, L.P. (Eds.), Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007. National Institute of Standards and Technology (NIST).

He, Y., Kayaalp, M., 2008. Biological Entity Recognition with Conditional Random Fields. AMIA. Annu. Symp. Proc. 2008, 293–297.

Hiemstra, D., Robertson, S., Zaragoza, H., 2004. Parsimonious Language Models for Information Retrieval, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04. ACM, New York, NY, USA, pp. 178–185. doi:10.1145/1008992.1009025

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., Milios, E.E., 2006. Information Retrieval by Semantic Similarity. Int J Semantic Web Inf Syst 2, 55–73.

Hoehndorf, R., Haendel, M., Stevens, R., Rebholz-Schuhmann, D., 2014. Thematic series on biomedical ontologies in JBMS: challenges and new directions. J. Biomed. Semant. 5, 15. doi:10.1186/2041-1480-5-15

Howe, J., 2008. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business, 1st ed. Crown Publishing Group, New York, NY, USA.

Huang, A., 2008. Similarity Measures for Text Document Clustering, in: Holland, J., Nicholas, A., Brignoli, D. (Eds.), . Presented at the New Zealand Computer Science Research Student Conference, pp. 49–56.

Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y., 2012. Improving Word Representations via Global Context and Multiple Word Prototypes, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 873–882.

Huang, M., Névéol, A., Lu, Z., 2011. Recommending MeSH terms for annotating biomedical articles. JAMIA 18, 660–667. doi:10.1136/amiajnl-2010-000055

Jaccard, P., 1912. The Distribution of the Flora in the Alpine Zone.1. New Phytol. 11, 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data Clustering: A Review. ACM Comput Surv 31, 264–323. doi:10.1145/331499.331504

Järvelin, K., Kekäläinen, J., 2002. Cumulated Gain-based Evaluation of IR Techniques. ACM Trans Inf Syst 20, 422–446. doi:10.1145/582415.582418

Jaynes, E.T., 1957. Information Theory and Statistical Mechanics. Phys. Rev. 106, 620–630. doi:10.1103/PhysRev.106.620

Jensen, M., Cox, A.P., Chaudhry, N., Ng, M., Sule, D., Duncan, W., Ray, P., Weinstock-Guttman, B., Smith, B., Ruttenberg, A., Szigeti, K., Diehl, A.D., 2013. The neurological disease ontology. J. Biomed. Semant. 4, 42. doi:10.1186/2041-1480-4-42

Jiang, J.J., Conrath, D.W., 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. ArXivcmp-Lg9709008.

Jiménez-Ruiz, E., Grau, B.C., Sattler, U., Schneider, T., Berlanga, R., 2008. Safe and economic re-use of ontologies: a logic-based methodology and tool support, in: Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications, ESWC'08. Springer-Verlag, Berlin, Heidelberg, pp. 185–199.

John, G., Langley, P., 1995. Estimating Continuous Distributions in Bayesian Classifiers, in: In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 338–345.

Spark Jones,, Walker, S., Robertson, S.E., 2000. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. Inf Process Manage 36, 779–808. doi:10.1016/S0306-4573(00)00015-7

Jupp, S., Bechhofer, S., Stevens, R., 2008. SKOS with OWL: Don't be Full-ish!, in: OWLED'08.

Kang, N., van Mulligen, E.M., Kors, J.A., 2011. Comparing and Combining Chunkers of Biomedical Text. J Biomed. Inform. 44, 354–360. doi:10.1016/j.jbi.2010.10.005

Katz, B., Uzuner, O., Yuret, D., 1998. Word Sense Disambiguation For Information Retrieval.

Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J., 2014. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014, in: Proceedings of CLEF 2014, Lecture Notes in Computer Science (LNCS). Springer.

Khan, L., Mcleod, D., Hovy, E., 2004. Retrieval Effectiveness of an Ontology-Based Model for Information Selection. VLDB J. 13, 71–85.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N., 2004. Introduction to the Bio-entity Recognition Task at JNLPBA, in: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications,

JNLPBA ’04. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 70–75.
Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D., 2004. Semantic annotation, indexing, and retrieval. J. Web Semant. 2, 49–79.
Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A., 2004. The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications, in: McIlraith, S.A., Plexousakis, D., Harmelen, F. van (Eds.), The Semantic Web – ISWC 2004, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 229–243.
Koehn, P., Hoang, H., Birch, A., Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E., 2007. Moses: open source toolkit for statistical machine translation. Presented at the Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, pp. 177–180.
Lafferty, J., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Morgan Kaufmann, pp. 282–289.
Lardilleux, A., Yvon, F., Lepage, Y., 2012. Hierarchical Sub-sentential Alignment with Anymalign, in: Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012). Trento, Italie, pp. 279–286.
Lassila, O., Swick, R.R., Wide, W., Consortium, W., 1998. Resource Description Framework (RDF) Model and Syntax Specification.
Leacock, C., Chodorow, M., 1998. Combining Local Context and WordNet Similarity for Word Sense Identification, in: WordNet: An Electronic Lexical Database. MIT Press.
Leaman, R., Gonzalez, G., 2008. BANNER: An executable survey of advances in biomedical named entity recognition, in: In Pac Symp Biocomput.
Lee, S., Lee, M., Kim, P., Jung, H., Sung, W.-K., 2010. OntoFrame S3: Semantic Web-Based Academic Research Information Portal Service Empowered by STAR-WIN, in: Aroyo, L., Antoniou, G., Hyvönen, E., Teije, A. ten, Stuckenschmidt, H., Cabral, L., Tudorache, T. (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 401–405.
Lenat, D.B., Guha, R.V., 1989. Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project, 1st ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
Letsche, T.A., Berry, M.W., 1997. Large-scale information retrieval with latent semantic indexing. Inf. Sci. 100, 105–137. doi:10.1016/S0020-0255(97)00044-3
Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals.
Liang, T., Shih, P.-K., 2005. Empirical Textual Mining to Protein Entities Recognition from PubMed Corpus, in: Montoyo, A., Muńoz, R., Métais, E. (Eds.), Natural Language Processing and Information Systems, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 56–66.
Lin, D., 1998. An Information-Theoretic Definition of Similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, ICML ’98. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 296–304.
Lindberg, D.A., Humphreys, B.L., McCray, A.T., 1993. The Unified Medical Language System. Methods Inf. Med. 32, 281–291.
Lin, H., Davis, J., 2010. Computational and Crowdsourcing Methods for Extracting Ontological Structure from Folksonomy, in: Aroyo, L., Antoniou, G., Hyvönen, E., Teije, A. ten, Stuckenschmidt, H., Cabral, L., Tudorache, T. (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 472–477.

Lin, H., Davis, J., Zhou, Y., 2010. Ontological Services Using Crowdsourcing. ACIS 2010 Proc.

Lin, J., Wilbur, W.J., 2007. PubMed related articles: a probabilistic topic-based model for content similarity. BMC Bioinformatics 8, 423. doi:10.1186/1471-2105-8-423

Liu, K., Wu, J., Peng, S., Zhai, C., Zhu, S., 2014. The Fudan-UIUC Participation in the BioASQ Challenge Task 2a: The Antinomyra system, in: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings. CEUR-WS.org, pp. 1311–1318.

Liu, S., Liu, F., Yu, C., Meng, W., 2004. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '04. ACM, New York, NY, USA, pp. 266–272. doi:10.1145/1008992.1009039

Liu, T.-Y., 2009. Learning to Rank for Information Retrieval. Found. Trends® Inf. Retr. 3, 225–331. doi:10.1561/1500000016

Lonsdale, D., Embley, D.W., Ding, Y., Xu, L., Hepp, M., 2010. Reusing Ontologies and Language Components for Ontology Generation. Data Knowl Eng 69, 318–330. doi:10.1016/j.datak.2009.08.003

Lopez, F., 1999. Overview Of Methodologies For Building Ontologies.

Lopez, F., Gómez-Pérez, A., Juristo, N., 1997. METHONTOLOGY: From ontological art towards ontological engineering, in: In AAAI-97 Spring Symposium on Ontological Engineering.

Lu, Z., Kim, W., Wilbur, W.J., 2009. Evaluation of Query Expansion Using MeSH in PubMed. Inf Retr 12, 69–80. doi:10.1007/s10791-008-9074-8

Maedche, A., Maedche, E., Staab, S., 2000. The TEXT-TO-ONTO Ontology Learning Environment, in: Software Demonstration at ICCS-2000 - Eight International Conference on Conceptual Structures.

Maedche, A., Maedche, E., Staab, S., Stojanovic, N., Sure, Y., Studer, R., 2001. SEmantic portAL - The SEAL approach, in: Spinning the Semantic Web. MIT Press, pp. 317–359.

Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R., 2003. An infrastructure for searching, reusing and evolving distributed ontologies, in: Proceedings of the 12th International Conference on World Wide Web. pp. 439–448.

Maedche, A., Staab, S., 2001. Ontology Learning for the Semantic Web. IEEE Intell. Syst. 16, 72–79. doi:10.1109/5254.920602

Mao, Y., Wei, C.-H., Lu, Z., 2014. NCBI at the 2014 BioASQ Challenge Task: Large-scale Biomedical Semantic Indexing and Question Answering, in: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings. CEUR-WS.org, pp. 1319–1327.

Mao, Y., Lu, Z., 2013. Ncbi at the 2013 bioasq challenge task: Learning to rank for automatic mesh indexing (Technical report).

McCandless, M., Hatcher, E., Gospodnetic, O., 2010. Lucene in Action, Second Edition: Covers Apache Lucene 3.0. Manning Publications Co., Greenwich, CT, USA.

McGuinness, D.L., Fikes, R., Hendler, J.A., Stein, L.A., 2002. DAML+OIL: An Ontology Language for the Semantic Web. IEEE Intell. Syst. 17, 72–80.

McInnes, B.T., Pedersen, T., 2013. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. J. Biomed. Inform., Special Section: Social Media Environments 46, 1116–1124. doi:10.1016/j.jbi.2013.08.008

McInnes, B.T., Pedersen, T., Pakhomov, S.V.S., 2009. UMLS-Interface and UMLS-Similarity : Open Source Software for Measuring Paths and Semantic Similarity. AMIA. Annu. Symp. Proc. 2009, 431–435.

Melamed, D., 1999. Bitext maps and alignment via pattern recognition. Comput. Linguist. 25, 107–130.

Meystre, S., Haug, P.J., 2006. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. J. Biomed. Inform. 39, 589–599. doi:10.1016/j.jbi.2005.11.004

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb. Med. Inform. 128–144.

Mihalcea, R., Moldovan, D., 2000. Semantic Indexing Using WordNet Senses, in: Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11, RANLPIR '00. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 35–45. doi:10.3115/1117755.1117760

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781.

Minard, A.-L., Ligozat, A.-L., Abacha, A.B., Bernhard, D., Cartoni, B., Deléger, L., Grau, B., Rosset, S., Zweigenbaum, P., Grouin, C., 2011. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. J. Am. Med. Inform. Assoc. amiajnl–2011–000154. doi:10.1136/amiajnl-2011-000154

Mizoguchi, R., 2003. Part 1: Introduction to Ontological Engineering. New Gen Comput 21, 365–384. doi:10.1007/BF03037311

Mizoguchi, R., 2004. Tutorial on ontological engineering Part 2: Ontology development, tools and languages. New Gener. Comput. 22, 61–96. doi:10.1007/BF03037281

Mizoguchi, R., Ikeda, M., Seta, K., Vanwelkenhuysen, J., 1995. Ontology for Modeling the World from Problem Solving Perspectives, in: Proc. of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing. pp. 1–12.

Morin, E., Prochasson, E., 2011. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora, in: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, BUCC '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 27–34.

Mork, J.G., Jimeno-Yepes, A., Aronson, A.R., 2013. The NLM Medical Text Indexer System for Indexing Biomedical Literature., in: Ngomo, A.-C.N., Paliouras, G. (Eds.), BioASQ@CLEF, CEUR Workshop Proceedings. CEUR-WS.org.

Mortensen, J., Alexander, P.R., Musen, M.A., Noy, N.F., 2013. Crowdsourcing Ontology Verification, in: Dumontier, M., Hoehndorf, R., Baker, C.J.O. (Eds.), Proceedings of the 4th International Conference on Biomedical Ontology, ICBO 2013, Montreal, Canada, July 7-12, 2013, CEUR Workshop Proceedings. CEUR-WS.org, pp. 40–45.

Mougin, F., Bodenreider, O., Burgun, A., 2009. Analyzing polysemous concepts from a clinical perspective: Application to auditing concept categorization in the UMLS. J. Biomed. Inform., Auditing of Terminologies 42, 440–451. doi:10.1016/j.jbi.2009.03.008

Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investig. 30, 3–26.

Névéol, A., Rogozan, A., Darmoni, S., 2006. Automatic indexing of online health resources for a French quality controlled gateway. Inf. Process. Manag. 42, 695–709. doi:10.1016/j.ipm.2005.01.003

Niles, I., Pease, A., 2001. Towards a Standard Upper Ontology, in: Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001, FOIS '01. ACM, New York, NY, USA, pp. 2–9. doi:10.1145/505168.505170

Nottelmann, H., Fuhr, N., 2003. From Retrieval Status Values to Probabilities of Relevance for Advanced IR Applications. Inf Retr 6, 363–388. doi:10.1023/A:1026080230789

Noy, N.F., Mcguinness, D.L., 2001. Ontology Development 101: A Guide to Creating Your First Ontology.

Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A., 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. gkp440. doi:10.1093/nar/gkp440

Och, F.J., Ney, H., 2003. A systematic comparison of various statistical alignment models. Comput Linguist 29, 19–51. doi:10.1162/089120103321337421

Okuda, T., Tanaka, E., Kasai, T., 1976. A Method for the Correction of Garbled Words Based on the Levenshtein Metric. IEEE Trans. Comput. C-25, 172–178. doi:10.1109/TC.1976.5009232

Papanikolaou, Y., Dimitriadis, D., Tsoumakas, G., Laliotis, M., Markantonatos, N., Vlahavas, I.P., 2014. Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine, in: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (Eds.), Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, CEUR Workshop Proceedings. CEUR-WS.org, pp. 1348–1360.

Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C.G., 2007. Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. 40, 288–299. doi:10.1016/j.jbi.2006.06.004

Peng, Y., Halper, M.H., Perl, Y., Geller, J., 2002. Auditing the UMLS for redundant classifications. Proc. AMIA Annu. Symp. AMIA Symp. 612–616.

Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M., Darmoni, S.J., 2009. F-MTI: outil d'indexation muititerminologique: application à l'indexation automatique de la SNOMED International, in: Risques, Technologies de l'Information pour les Pratiques Médicales, Informatique et Santé. Springer Paris, pp. 57–68.

Ponte, J.M., Croft, W.B., 1998. A Language Modeling Approach to Information Retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98. ACM, New York, NY, USA, pp. 275–281. doi:10.1145/290941.291008

Porter, M., 1980. An algorithm for suffix stripping. Program 14, 130–137. doi:10.1108/eb046814

Pradhan, S., Elhadad, N., South, B.R., Martínez, D., Christensen, L.M., Vogel, A., Suominen, H., Chapman, W.W., Savova, G.K., 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013, in: Forner, P., Navigli, R., Tufis, D., Ferro, N. (Eds.), Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013, CEUR Workshop Proceedings. CEUR-WS.org.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257–286. doi:10.1109/5.18626

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. 19, 17–30. doi:10.1109/21.24528

Rajpathak, D., Motta, E., Roy, R., 2001. A generic task ontology for scheduling applications. Presented at the International Conference on Artificial Intelligence (IC AI'2001), Las Vegas, USA.

Renoust, B., Melançon, G., Viaud, M.-L., 2013. Measuring Group Cohesion in Document Collections 373–380. doi:10.1109/WI-IAT.2013.53

Ren, Z., Lü, Y., Cao, J., Liu, Q., Huang, Y., 2009. Improving statistical machine translation using domain bilingual multiword expressions, in: Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, MWE '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 47–54.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.

Reynar, J.C., Ratnaparkhi, A., 1997. A Maximum Entropy Approach to Identifying Sentence Boundaries, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, ANLC '97. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 16–19. doi:10.3115/974557.974561

Ribadas-Pena, F.J., Ibañez, L.M. de C., Bilbao, V.M.D., Romero, A.E., 2013. Two Hierarchical Text Categorization Approaches for BioASQ Semantic Indexing Challenge, in: Ngomo, A.-C.N., Paliouras, G. (Eds.), Proceedings of the First Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013, CEUR Workshop Proceedings. CEUR-WS.org.

Rijsbergen, C.J.V., 1979. Information Retrieval, 2nd ed. Butterworth-Heinemann, Newton, MA, USA.

Robertson, S.E., Jones, K.S., 1976. Relevance weighting of search terms. J. Am. Soc. Inf. Sci. 27, 129–146. doi:10.1002/asi.4630270302

Robertson, S.E., van Rijsbergen, C.J., Porter, M.F., 1981. Probabilistic Models of Indexing and Searching, in: Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval, SIGIR '80. Butterworth & Co., Kent, UK, UK, pp. 35–56.

Robertson, S.E., Walker, S., 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, pp. 232–241.

Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., 1996. Okapi at TREC-3. pp. 109–126.

Roche, C., 2005. Terminologie et ontologie. Langages 157, 48–62. doi:10.3917/lang.157.0048

Roche, C., 2012. Ontoterminology: How to unify terminology and ontology into a single paradigm, in: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012. European Language Resources Association (ELRA), pp. 2626–2630.

Rosse, C., Mejino Jr., J.L.V., 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J. Biomed. Inform., Unified Medical Language System 36, 478–500. doi:10.1016/j.jbi.2003.11.007

Ruch, P., 2006. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 22, 658–664. doi:10.1093/bioinformatics/bti783
Salton, G., Fox, E.A., Wu, H., 1983. Extended Boolean Information Retrieval. Commun ACM 26, 1022–1036. doi:10.1145/182.358466
Salton, G., McGill, M.J., 1986. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA.
Salton, G., Wong, A., Yang, C.S., 1975. A Vector Space Model for Automatic Indexing. Commun ACM 18, 613–620. doi:10.1145/361219.361220
Sánchez, D., Batet, M., 2011. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. J. Biomed. Inform. 44, 749–759. doi:10.1016/j.jbi.2011.03.013
Sanderson, M., 1994. Word Sense Disambiguation and Information Retrieval, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, pp. 142–151.
Sarasua, C., Simperl, E., Noy, N.F., 2012. CrowdMap: Crowdsourcing Ontology Alignment with Microtasks, in: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (Eds.), The Semantic Web – ISWC 2012, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 525–541.
Schadow, G., McDonald, C.J., 2003. Extracting Structured Information from Free Text Pathology Reports. AMIA. Annu. Symp. Proc. 2003, 584–588.
Scheuermann, R.H., Ceusters, W., Smith, B., 2009. Toward an Ontological Treatment of Disease and Diagnosis. Summit Transl. Bioinforma. 2009, 116–120.
Schmid, H., 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees.
Settles, B., 2004. Biomedical named entity recognition using conditional random fields and rich feature sets, in: In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications (NLPBA. pp. 104–107.
Simperl, E., 2009. Reusing ontologies on the Semantic Web: A feasibility study. Data Knowl Eng 68, 905–925. doi:10.1016/j.datak.2009.02.002
Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255. doi:10.1038/nbt1346
Soergel, D., Lauser, B., Liang, A.C., Fisseha, F., Keizer, J., Katz, S., 2004. Reengineering Thesauri for New Applications: The AGROVOC Example. J Digit Inf 4.
Soldatova, L.N., King, R.D., 2005. Are the current ontologies in biology good ontologies? Nat. Biotechnol. 23, 1095–1098. doi:10.1038/nbt0905-1095
Sorg, P., Cimiano, P., 2012. Exploiting Wikipedia for cross-lingual and multilingual information retrieval. Data Knowl. Eng., Applications of Natural Language to Information Systems 74, 26–45. doi:10.1016/j.datak.2012.02.003
Soukoreff, R.W., MacKenzie, I.S., 2001. Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic, in: CHI '01 Extended Abstracts on Human Factors in Computing Systems, CHI EA '01. ACM, New York, NY, USA, pp. 319–320. doi:10.1145/634067.634256
Sowa, J.F., 2000. Knowledge Representation: Logical, Philosophical and Computational Foundations. Brooks/Cole Publishing Co., Pacific Grove, CA, USA.
Spackman, K.A., Reynoso, G., 2004. Examining SNOMED from the perspective of formal ontological principles: Some preliminary analysis and observations, in: Hahn, U.

(Ed.), KR-MED 2004, First International Workshop on Formal Biomedical Knowledge Representation, Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation, Whistler, BC, Canada, 1 June 2004, CEUR Workshop Proceedings. CEUR-WS.org, pp. 72–80.

Spyromitros, E., Tsoumakas, G., Vlahavas, I., 2008. An Empirical Study of Lazy Multilabel Classification Algorithms, in: Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications, SETN '08. Springer-Verlag, Berlin, Heidelberg, pp. 401–406. doi:10.1007/978-3-540-87881-0_40

Stokes, N., Li, Y., Cavedon, L., Zobel, J., 2009. Exploring Criteria for Successful Query Expansion in the Genomic Domain. Inf Retr 12, 17–50. doi:10.1007/s10791-008-9073-9

Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge Engineering: Principles and Methods. Data Knowl Eng 25, 161–197. doi:10.1016/S0169-023X(97)00056-6

Suarez-Figueroa, M., Gomez-Perez, A., 2009. NeOn Methodology for Building Ontology Networks: a Scenario-based Methodology, in: on SOFTWARE, S.& S.T.S. 2009 I.C. (Ed.), . Presented at the Proceedings of the International Conference on SOFTWARE, SERVICES & SEMANTIC TECHNOLOGIES (S3T 2009).

Subramaniam, L.V., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V.S., Kamesam, P.V., Kothari, R., 2003. Information Extraction from Biomedical Literature: Methodology, Evaluation and an Application, in: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03. ACM, New York, NY, USA, pp. 410–417. doi:10.1145/956863.956941

Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G.K., Elhadad, N., Pradhan, S., South, B.R., Mowery, D., Jones, G.J.F., Leveling, J., Kelly, L., Goeuriot, L., Martínez, D., Zuccon, G., 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013, in: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, Lecture Notes in Computer Science. Springer, pp. 212–231. doi:10.1007/978-3-642-40802-1_24

Suominen, O., Viljanen, K., HyvÄnen, E., 2007. User-Centric Faceted Search for Semantic Portals, in: Franconi, E., Kifer, M., May, W. (Eds.), The Semantic Web: Research and Applications, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 356–370.

Sure, Y., Staab, S., Studer, R., Gmbh, O., 2003. On-To-Knowledge Methodology (OTKM), in: Handbook on Ontologies, International Handbooks on Information Systems. Springer, pp. 117–132.

Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.J., 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics 6, S3. doi:10.1186/1471-2105-6-S1-S3

Tayeb, M., Julien, G., Hocine, A., Michel, J., Stefan, D., 2011. Automatic methods for mapping Biomedical terminologies in a Health Multi-Terminology Portal, in: EGC 2011 : Atelier Extraction de Connaissances et Santé, 2011.

Toutanova, K., Klein, D., Manning, C.D., Singer, Y., 2003. Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 173–180. doi:10.3115/1073445.1073478

Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., Rebholz-Schuhmann, D., 2009. MeSH Up: effective MeSH text classification for improved document retrieval. Bioinforma. Oxf. Engl. 25, 1412–1418. doi:10.1093/bioinformatics/btp249

Tsatsaronis, G., Panagiotopoulou, V., 2009. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, EACL '09. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 70–78.

Tsatsaronis, G., Schroeder, M., Paliouras, G., Almirantis, Y., Androutsopoulos, I., Gaussier, É., Gallinari, P., Artières, T., Alvers, M.R., Zschunke, M., Ngomo, A.-C.N., 2012. BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012, AAAI Technical Report. AAAI.

Tsoumakas, G., Katakis, I., Vlahavas, I., 2010. Mining multi-label data, in: In Data Mining and Knowledge Discovery Handbook. pp. 667–685.

Tudorache, T., Nyulas, C., Noy, N.F., Musen, M.A., 2013. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the Web. Semant Web 4, 89–99.

Uschold, M., 1998. Knowledge Level Modelling: Concepts and Terminology. Knowl Eng Rev 13, 5–29. doi:10.1017/S0269888998001040

Uschold, M., Gruninger, M., 1996. Ontologies: Principles, methods and applications. Knowl. Eng. Rev. 11, 93–136.

Uschold, M., King, M., 1995. Towards a Methodology for Building Ontologies, in: In Workshop on Basic Ontological Issues in Knowledge Sharing, Held in Conjunction with IJCAI-95.

Uzuner, O., Katz, B., Yuret, D., 1999. Word Sense Disambiguation for Information Retrieval, in: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, AAAI '99/IAAI '99. American Association for Artificial Intelligence, Menlo Park, CA, USA, p. 985–.

Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L., 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J. Am. Med. Inform. Assoc. 18, 552–556. doi:10.1136/amiajnl-2011-000203

Velardi, P., Fabriani, P., Missikoff, M., 2001. Using Text Processing Techniques to Automatically Enrich a Domain Ontology, in: Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001, FOIS '01. ACM, New York, NY, USA, pp. 270–284. doi:10.1145/505168.505194

Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F., 2006. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies, in: Buitelaar, P., Cimiano, P., Magnini, B. (Eds.), Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press.

Villazón-Terrazas, B., Gómez-Pérez, A., 2012. Reusing and Re-engineering Non-ontological Resources for Building Ontologies, in: Suárez-Figueroa, M. del C., Gómez-Pérez, A., Motta, E., Gangemi, A. (Eds.), Ontology Engineering in a Networked World. Springer, pp. 107–145.

Voorhees, E.M., 1993. Using WordNet to disambiguate word senses for text retrieval, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '93. ACM, New York, NY, USA, pp. 171–180. doi:10.1145/160688.160715

Voorhees, E.M., 1994. Query expansion using lexical-semantic relations, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '94. Springer-Verlag New York, Inc., New York, NY, USA, pp. 61–69.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hübner, S., 2001. Ontology-Based Integration of Information - A Survey of Existing Approaches. pp. 108–117.

Wang, X., 2007. Rule-based protein term identification with help from automatic species tagging, in: In Proceedings of CICLING 2007. pp. 288–298.

Widdows, D., Peters, S., Cederberg, S., Chan, C.-K., Steffen, D., Buitelaar, P., 2003. Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using UMLS, in: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13, BioMed '03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 9–16. doi:10.3115/1118958.1118960

Wong, S.K.M., Ziarko, W., Wong, P.C.N., 1985. Generalized Vector Spaces Model in Information Retrieval, in: Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '85. ACM, New York, NY, USA, pp. 18–25. doi:10.1145/253495.253506

Wu, Z., Palmer, M., 1994. Verbs Semantics and Lexical Selection, in: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, ACL '94. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 133–138. doi:10.3115/981732.981751

Xu, Y., Hong, K., Tsujii, J., Chang, E.I.-C., 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. JAMIA 19, 824–832. doi:10.1136/amiajnl-2011-000776

Zhang, L., Yu, Y., Yang, Y., Zhou, J., Lin, C., 2005. An Enhanced Model for Searching in Semantic Portals, in: In WWW '05: Proceedings of the 14th International Conference on World Wide Web. ACM Press, pp. 453–462.

Zhang, M.-L., Zhou, Z.-H., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognit. 40, 2038–2048. doi:10.1016/j.patcog.2006.12.019

Zhang, S., Elhadad, N., 2013. Unsupervised Biomedical Named Entity Recognition: Experiments with Clinical and Biological Texts. J Biomed. Inform. 46, 1088–1098. doi:10.1016/j.jbi.2013.08.004

Zhou, W., Yu, C., Smalheiser, N., Torvik, V., Hong, J., 2007. Knowledge-intensive Conceptual Retrieval and Passage Extraction of Biomedical Literature, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07. ACM, New York, NY, USA, pp. 655–662. doi:10.1145/1277741.1277853

Zhou, X., Zhang, X., Hu, X., 2006a. MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup, in: Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence, PRICAI'06. Springer-Verlag, Berlin, Heidelberg, pp. 1145–1149.

Zhou, X., Zhang, X., Hu, X., 2006b. Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR, in: Proceedings of the 28th European Conference on Advances in Information Retrieval, ECIR'06. Springer-Verlag, Berlin, Heidelberg, pp. 444–455. doi:10.1007/11735106_39

Zhu, D., Li, D., Carterette, B., Liu, H., 2013. An Incremental Approach for MEDLINE MeSH Indexing, in: Ngomo, A.-C.N., Paliouras, G. (Eds.), Proceedings of the First

Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013, CEUR Workshop Proceedings. CEUR-WS.org.

Zouaq, A., Nkambou, R., 2010. A Survey of Domain Ontology Engineering: Methods and Tools, in: Nkambou, R., Bourdeau, J., Mizoguchi, R. (Eds.), Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence. Springer Berlin Heidelberg, pp. 103–119.

Zweigenbaum, P., 2004. L'UMLS entre langue et ontologie: une approche pragmatique dans le domaine médical. Rev. Intell. Artif. 18, 111–137. doi:10.3166/ria.18.111-137

Zweigenbaum, P., Jacquemart, P., Grabar, N., Habert, B., 2001. Building a text corpus for representing the variety of medical language. Stud. Health Technol. Inform. 290–294.